Group sequential tests with random information accrual

Barber, Stuart

Award date:
1999

Awarding institution:
University of Bath

Group Sequential Tests With Random Information Accrual

submitted by

Stuart Barber

for the degree of

PhD in Statistics

of the

University of Bath

1999

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Group Sequential Tests With Random Information Accrual

submitted by

Stuart Barber

for the degree of PhD

of the

University of Bath

1999

Summary

Following the work of Pocock (1977) and O'Brien & Fleming (1979), group sequential methods have become commonly used in the monitoring of clinical trials. An appealing property of these methods is the possibility of terminating the trial once there is sufficient evidence in favour of one of the treatments being studied. This can lead to great savings in time and resources and, more importantly, reduce the number of patients who receive the inferior treatment. Work by Eales & Jennison (1992), based on earlier research by Lai (1973), has shown that group sequential designs which minimise the expected sample size can be found when the type I and type II error rates are equal. However, these optimal tests lack flexibility in that the precise schedule of analyses must be fixed before the commencement of the trial.

We generalise the method of Eales & Jennison to cope with unequal error rates and show that commonly used and popular existing methods can be reasonably efficient in the sense of the expected sample size. However, there are significant improvements in expected sample size to be gained from using the optimal tests.

A popular existing method is the error spending approach, proposed by Lan & DeMets (1983). Not only can this method be reasonably efficient in terms of expected sample size, but it has great flexibility in dealing with trials where the observations accrue at an unanticipated rate. We develop optimal designs which do not require a fixed schedule of analyses and use these optimal design to assess the strengths and weaknesses of the error spending approach.

Acknowledgements

Many people deserve my gratitude for their help and support over the past three years. In particular, I wish to thank my supervisor, Professor Christopher Jennison, for his advice and guidance.

My family have been a constant source of support for the last three years, for which I am indebted to them.

My special thanks go to Siân Blake for her support and for having the patience to read the non-statistical aspects of this work.

I am also grateful to the Ph.D. students of 1W4.19 and my other friends in Bath for making the past three years so enjoyable.

Finally, I am grateful to the EPSRC for their financial support over the past three years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Outline

Sequential methods are of great potential use in clinical trials. By allowing the possibility of stopping the trial at an early stage if there is sufficient evidence in favour of one of the treatments on trial, significant saving in time and resources can be made. Moreover, it is possible to greatly reduce the expected number of patients involved in the trial and thus minimise the number of patients who will be given an inferior treatment. We shall discuss a means of maximising the benefits of sequential clinical trials in a number of settings.

In this chapter we discuss the clinical trials setting and the development of sequential methods, with some comments on the practical implementation of sequential clinical trials. Chapter 2 introduces two rich families of sequential designs which are commonly used. Chapters 3 to 6 discuss the optimisation of sequential designs in several settings and use the optimal designs to assess the performance of the existing designs introduced in chapter 2. We conclude with some conclusions and recommendations for designing sequential trials in chapter 7, along with some possible avenues for future research.

1.2 Why sequential clinical trials?

In this section, the background to sequential clinical trials is discussed, with the motivation for such trials and the practical considerations that lead to the use of group sequential trials described in §1.2.2. One criterion for optimality of such trials, minimising the expected number of patients on trial, is defined in §1.2.3. Our goal is to find optimal group sequential tests with respect to this criterion in a number of circumstances and to use these optimal tests to evaluate the performance of some existing and commonly used designs. Two such existing methods are described in chapter 2 and their performance with respect to our optimality criteria is assessed in chapters 3 to 6. Some practical aspects of implementing group sequential clinical trials are also discussed in §1.3. Firstly, a brief description of clinical trials is given in §1.2.1. More details regarding the design and statistical analysis of non-sequential clinical trials are given in numerous books including Pocock (1983), while the use of sequential designs in clinical trials is discussed by Whitehead (1992) and Jennison & Turnbull (2000).

1.2.1 Clinical trials

The goal of a clinical trial is to determine the efficacy of one or more treatments for some medical condition, relative to a control or each other. Pocock defines a clinical trial as "any form of planned experiment which involves patients and is designed to elucidate the most appropriate treatment of future patients with a given medical condition" (Pocock, 1983, chapter 1). In clinical trials involving human subjects the ethics and conduct of the trial must be more closely scrutinised and regulated than, for example, early-stage pharmaceutical trials on laboratory animals.

We assume that a new treatment is being compared to a control, which could be either a placebo or an established treatment. We shall assume that the difference between

the treatment on trial and the control can be measured in some numerical fashion, and shall call the treatment difference μ, with positive treatment difference ($\mu > 0$) implying superiority of the new treatment. The available data consist of observations $X_1, X_2, \ldots$, where each $X_i \sim N(\mu, \sigma^2)$ independently with σ^2 known. For example, these observations could be the differences between matched pairs of individuals, or the difference in response to the treatments for one individual.

In order to address the question "is the new treatment superior to the control?", we shall wish to test the null hypothesis H_0:$\mu \leq 0$ against the one-sided alternative H_1:$\mu > 0$ with some chosen size α at $\mu = 0$ and with power $1 - \beta$ fixed at $\mu = \delta$. Both one- and two-sided tests are justified in different situations; a brief discussion of this is given in Pampallona & Tsiatis (1994), and the methods that we shall discuss are easily adapted to the two-sided case. In the trial design stage, δ would be chosen to be some clinically significant and plausible treatment effect. The null hypothesis represents the situation where the new treatment is not superior to the control and the alternative corresponds to the situation where the new treatment is sufficiently superior to the control to make its use desirable. Note that the null hypothesis incorporates two situations; where the new treatment is equivalent to the control and when it is inferior. If the control is an existing treatment, it is irrelevant whether a new treatment is equivalent or inferior to it as in either case the new treatment will not be used. Similarly, if the control is a placebo, finding the new treatment to be equivalent to the control implies that the only benefit gained from the new treatment is simply a psychological placebo effect, so again the new treatment will not be put into use.

Initially, the assumption of independent and identically distributed normal data appears to pose an unrealistically simple problem of little interest. However, in the group sequential setting, there are many applications where the statistical formulation has, at least asymptotically, the same distribution as in this case. For examples of the wide applicability of this distribution see Jennison & Turnbull (1997). Jennison

& Turnbull show that this asymptotic distribution occurs for sequences of maximum likelihood estimators of parameters in statistical models. They specifically discuss parameters of a generalised linear model and a Cox proportional hazards model. Whitehead (1992) discusses a framework which can be used for problems involving binary, Poisson and ordinal responses and with censored survival data, although without explicitly considering the joint distribution of the summary statistics involved. More generally, Scharfstein, Tsiatis & Robins (1997) show that the relevant joint distribution occurs when a single parameter of a parametric or semi-parametric model is efficiently estimated in a sequential fashion.

In more complicated examples than that of simple independent and identically distributed normal observations, the sample size n is no longer directly interpreted as the number of observations, but as the information level; in most cases this will be Fisher's information. This implies a "time" scale where we start the trial with zero information and accumulate data until a conclusion is reached; this approach is often referred to as a maximum information design. This information level is distinct from calendar time as the two may proceed at different rates. For example, if the trial is examining the survival rates of patients with some terminal condition, there may be an initial period of recruitment to the trial after which a lengthy follow-up period will take place. The information available is directly proportional to the number of deaths which have occurred; the rate of deaths will not be constant over time. Thus, the information level will have accrued at a rate which will not be proportional to calendar time. A more detailed discussion of the differences and links between the passing of calendar time and accrual of information is given by Lan, Reboussin & DeMets (1994).

1.2.2 Sequential and group sequential tests

In large clinical trials, which may involve thousands of patients, it is usual to monitor accumulating data for a number of reasons. For example, there is always the

possibility of unanticipated toxicity of a treatment, or of other undesirable side-effects, necessitating the early termination of the trial. Many regulatory bodies now require the monitoring of data as the trial progresses. One example is found in the Food and Drug Administration's guidelines for clinical trials, which states that "it is recognized that safety must be monitored in all trials; therefore, the need for formal procedures to cover early stopping for safety reasons should always be considered" (FDA, 1998).

A natural consequence of this monitoring is a desire to stop the trial before the planned number of observations have been seen if a clear difference in efficacy between the two treatments is observed earlier in the trial. To this end, sequential and group sequential methods have been developed.

Initial work on sequential schemes was based on industrial quality control problems. Wald (1945) proposed the sequential probability ratio test, which assumes that there is no maximum to the number of observations that can be taken. Later work by Anderson (1960) looked at the sequential analysis of data when a fixed maximum number of observations can be taken. All these proposals were based on analysing the data after each observation was made. A more practical proposal for use in the medical setting was made by Pocock (1977), where data are analysed in groups, referred to as group sequential analysis. This made the logistics of interim analyses practical in a medical context. We shall refer to methods requiring analysis of the data after every observation as fully sequential methods. The term sequential analysis will refer to both fully sequential and group sequential methods. A more detailed discussion of the history of sequential analysis is given by Ghosh (1991).

In order to carry out the hypothesis test described in §1.2.1, we assume that there will be a maximum of K interim analyses of the data, with the trial terminating at analysis K if it has not done so at an earlier analysis. Note that we refer to all analyses of the data, including the final possible analysis, as interim analyses to distinguish them from the analysis taken at the conclusion of a non-sequential trial. Define n_i to be the total

number of observations taken by analysis i. We note that the maximum sample size n_K will be larger than the sample size required for a fixed sample size test allowing no interim analyses of the same hypotheses with equal size and power. Initially, in chapters 3 and 4, we only consider problems where K and the sample sizes $n_1, \ldots, n_K$ are fixed as part of the trial design. In chapter 5, we shall go on to consider what happens when the sequence of sample sizes is not fixed in advance and in chapter 6 we relax the requirement of fixing the maximum number of analyses.

Commonly, sequential methods are represented by the plotting of an information statistic (in the case of independent and identically distributed normal data we have discussed above the information statistic is simply the cumulative sample size) against a summary statistic of the observed data (such as the sum of the data observed to date). At each analysis there will be a continuation region, and the trial will be terminated with either the rejection or acceptance of H_0 if the observed summary statistic is outside the continuation region. The bounds of the continuation regions at each analysis are said to form the sequential boundary and the test will be terminated once the test boundary is crossed. Many different rules exist to determine the continuation region at each analysis, and hence the sequential boundary; two popular and commonly used rules are described in chapter 2. In chapters 3 to 6, we describe methods for finding optimal group sequential boundaries, but first we must define our optimality criteria, which we discuss in the next section.

1.2.3 Optimal group sequential tests

One of the main motivations for group sequential tests is the ethical imperative to stop the trial with a recommendation as to the superior treatment as early as possible; the aim being to reduce the number of patients randomised to the inferior treatment. Work has been done on adapting the allocation of new patients to treatments in light of the data so far seen in a trial, but that is beyond the scope of our current work; for

a review of this subject, see Basu, Box & Ghosh (1991) and for more recent work see Coad & Rosenberger (1999) or Jennison & Turnbull (2000, chapter 17). Stopping the trial as early as possible also reduces the cost of the trial, in terms of both financial and other resources which is an attraction to pharmaceutical companies. We shall take the desire to minimise the expected number of observations as our prime motivation in designing group sequential tests.

Note that the sample size of the test (i.e. the number of observations taken) is a random variable, as opposed to the fixed sample size tests used when there are no interim analyses. Define N to be the total number of observations taken and n_i to be the cumulative sample size at analysis i. Then N takes values in the set $\{n_1 \ldots, n_K\}$ and $\mathbb{P}_\mu\{N = n_i\} = \mathbb{P}_\mu\{\text{trial terminates at analysis } i\}$. We shall describe a test as optimal if it minimises $\mathbb{E}_\mu\{N\}$ for a particular value or set of values of μ; these expectations are referred to as objective functions. For example, one such objective function would be $\mathbb{E}_0\{N\}$, the expected sample size if, in truth, the new treatment is equal in efficacy to the control. In §3.1.1 we define several objective functions, for different values of μ or averaged over several values of μ.

Having defined objective functions and set our initial goal to be the minimising of these, we note that in some cases it will not be appropriate to seek the minimising of expected sample size as the primary criterion for test design. For example, in a clinical trial where the treatments are of similar cost, toxicity and efficacy, information on long-term effects may be desired. Alternatively, it is possible to imagine a situation where there is initial superiority for one treatment, but where survival for the two treatment arms converges or even crosses as follow-up increases. Also, as we shall see, some tests which have extremely low expected sample sizes have the undesirable property of having a very large maximum sample size. Even if the probability of reaching the maximum sample size is small, a design which allows the possibility of taking twice as many observations as the equivalent fixed sample size design will be unattractive.

1.3 Implementing group sequential clinical trials

Having briefly discussed the motivation for and background to group sequential tests, it is important to remember that any group sequential design must be practical to implement if it is to be used in a clinical trial. It is this criterion which has lead to the development of group sequential designs in place of fully sequential methods; the logistics of analysing the accumulating data after each observation are prohibitive.

Commonly, information collected in the course of a large scale clinical trial will be periodically reviewed by a data monitoring committee which will have the responsibility of recommending the continuation or termination of the trial to the investigators. In §1.3.2, the role of the data monitoring committee is considered, but first we discuss several design criteria for group sequential trials.

1.3.1 Design criteria

Many possible schemes have been proposed for designing and carrying out a group sequential clinical trial. By "scheme", we refer to the rules used to determine when the interim analyses will occur and what action will be taken as a result of the observed data at these analyses. Before selecting a scheme, it is important to consider the reason for using a group sequential design and to note certain points which are valid irrespective of the scheme used.

As with any statistical analysis, the endpoint(s) of interest must be identified. While the vast majority of work on group sequential tests has been devoted to trials with a single endpoint, the question of multivariate endpoints has been addressed. Jennison & Turnbull (1993) considered bivariate responses, monitoring both the efficacy and safety of a new treatment. Cook & Farewell (1996) consider combining the clinical endpoint of a trial and a surrogate endpoint to form a single test statistic. There have also been ad-hoc treatments of this problem, such as example 2 of Geller & Pocock (1987).

Geller & Pocock go on to discuss this problem briefly, making three suggestions. One is to combine endpoints of interest, so for example recurrence of a condition or death could be combined into one event. A second is to test each endpoint separately at each interim analysis, making suitable modifications to avoid error inflation. Thirdly, it may be possible to combine multiple endpoints into a global test statistic.

The number and timing of interim analyses planned must be considered. In part, the selection of a group sequential scheme will influence this choice. However, for all schemes the increase in efficiency possible through group sequential testing is mostly gained with a small number of analyses, as shown by Pocock (1982), Jennison (1987), Eales & Jennison (1992) and §3.2.1 of this thesis, amongst others.

Moreover, while some schemes such as the error-spending approach proposed by Lan & DeMets (1983) (see §2.3) or Whitehead's triangular test (Whitehead, 1992, chapter 4), have the advantage of flexibility of the timing of analyses, this flexibility must not be abused. The choice of analysis times should not be response-dependent, although altering the times of analysis on the basis of unanticipated accrual rates is acceptable. By fixing the analysis times as part of the trial design, accusations of such manipulation of the trial can be avoided. Another reason for planning analysis times in advance is the purely practical point that members of the data monitoring committee will be able to schedule meetings to discuss the trial.

It is also important to note that the analyses do not need to occur at equally spaced intervals of calendar time or after equal increments in information level, although some group sequential schemes do stipulate analyses at equally spaced intervals, usually equal information level increments. If the trial is investigating treatments for a condition with a delay between treatment and onset of effect, or with long follow-up, it may be desirable to delay the initial analyses until the treatment has had time to take effect.

Potentially the most important choice to be made in group sequential trial design is the

choice of which scheme to use. This question should be approached by first considering the reasons for monitoring the data. If the data are only subjected to interim analysis for safety reasons and to comply with regulatory body guidelines, with little or no expectation of early termination, a scheme which is conservative at early analyses will be appropriate. The final analysis of such a scheme will be similar to an equivalent non-sequential design, and as such will be easier to interpret. Conversely, a trial where there is believed to be the possibility for a large treatment effect has greater potential for early termination and in such cases a scheme which is less conservative at early analyses will be more appropriate.

Another question which will influence the choice of scheme is the rate of arrival of the data. Many schemes assume that the number of observations available at each analysis will be fixed in advance. While such schemes are usually robust to small departures from this assumption, as shown by Pocock (1977), gross departures can lead to problems as described by Proschan, Follman & Waclawiw (1992). In a trial where fears exist that the rate of data arrival might be unpredictable or erratic, a flexible approach such as error spending (Lan & DeMets, 1983) or Whitehead's triangular test (Whitehead, 1992, chapter 4) might be considered.

1.3.2 The role of the data monitoring committee

It is now common practice for the progress of a large scale clinical trial to be monitored by a data monitoring committee (also sometimes referred to as a data and safety monitoring board, or other equivalent terms). The data monitoring committee will either be internal to the institution carrying out the trial or an independent body. In either case, the committee has several duties. One of these is to ensure the trial protocol is adhered to. The committee may also find it appropriate to recommend alterations to the protocol; for example, if practical circumstances dictate a change in treatment regime or inclusion criteria, or if new information becomes available from other trials.

Whenever feasible, a clinical trial will be blinded; that is the patients and clinicians will not know which treatment is being given to each individual. The goal of this is to ensure that there is no conscious or unconscious bias and to make the placebo effect apply equally to both control and treatment groups. Frequently, the interim analyses involved in a sequential scheme will require the breaking of this blind, at least by the statistician analysing the data, but this is not always the case. In some situations it would be possible to report that the trial was nearing the boundary indicating superiority of treatment A without saying whether treatment A was the new or control treatment. Even if the blind is not broken, the data monitoring committee should ideally be comprised of individuals not directly involved with the trial and with no possible conflict of interest.

A potential reason for the termination of the trial is that sufficient evidence has been accrued to support the superiority of one of the treatments (or to support the hypothesis that there is no treatment difference). This is precisely the function of sequential monitoring schemes, and hence one role of the committee is to act as custodians of the scheme. The data monitoring committee may also recommend the termination of the trial under circumstances other than the crossing of the sequential boundary. For example, if there is unanticipated mortality or other negative side effects on one treatment it may be appropriate to terminate the trial.

As previously stated, there may be alternative reasons for terminating the trial before the planned end other than the sequential boundary being crossed. Conversely, there may also be a desire to continue the trial after the sequential boundary has been crossed. For example, in an equivalence trial the test may have concluded that there are no differences between the two treatments. If there are no known side effects and the costs of the treatments are similar, then the data monitoring committee may recommend the continuation of the trial to obtain further information on the long-term effects of the new treatment.

Thus, the stopping rule determined by the group sequential scheme may be interpreted as a somewhat flexible guideline rather than a rigid rule. While this may be statistically and even ethically dubious, given the complexity of the decision process involved in monitoring the trial, it is important that the data monitoring committee retains the ability to recommend actions contrary to those dictated by the stopping rule. However, it must be noted that continuing the trial when the scheme dictates the termination of the trial invalidates much of the existing methodology for post-trial analysis. More details of this are given in the next section.

Further discussion of the role and responsibilities of the data monitoring committee can be found in Korn & Simon (1996) or Fleming & DeMets (1993).

1.3.3 Post-trial estimation and analysis

On conclusion of a clinical trial, whether or not the trial is sequential in nature, more detailed statistical analysis will be required than a simple statement that the new treatment studied is or is not superior in efficacy to the control at some pre-specified level of significance. The treatment effect must be estimated, either as a point estimate or preferably a confidence interval; p-values will frequently be required as a measure of evidence by regulatory bodies and clinicians who must decide whether or not to use the new treatment.

Conventional fixed-sample methods for carrying out statistical analyses are invalid after completion of a sequential trial and alternatives must be used. Most frequentist methods are based on an ordering of the possible sample space of observed data and at what point the trial was terminated (earlier termination of the trial often being taken as more conclusive evidence of treatment difference), such as that described by Jennison & Turnbull (1991). Post trial analysis is discussed in depth by Whitehead (1992, chapter 5) and by Jennison & Turnbull (2000, chapter 8).

Chapter 2

Two existing group sequential test designs

2.1 Introduction

In this chapter, two existing designs for group sequential tests are introduced and discussed. These designs produce rich families of tests, allowing a wide choice of tests in each family. The first of these two designs is the Δ-family of one-sided tests due to Pampallona & Tsiatis (1994), following work by Emerson & Fleming (1989), discussed in §2.2. In §2.3 we introduce the error-spending method of Lan & DeMets (1983), first in the two-sided context where this method was initially proposed and then in the case of one-sided tests. These methods are assessed in terms of maximum and expected sample size in various situations in the following chapters.

2.2 The Δ-family

Many authors have proposed group sequential schemes based on a fixed series of analysis times. The first group sequential designs, by Pocock (1977) and O'Brien & Fleming (1979), were two-sided test designs of this form. Wang & Tsiatis (1987) proposed a family of two-sided tests which included the Pocock and O'Brien & Fleming tests as special cases. This rich family of tests was extended by Emerson & Fleming (1989) and by Pampallona & Tsiatis (1994) to include two-sided tests with the possibility of stopping in favour of the null hypothesis and to include one-sided tests. We now describe this family of tests, which we index by the parameter Δ. We shall refer to this family of tests as the Δ-family.

Suppose we wish to test H_0:$\mu \leq 0$ against H_1:$\mu > 0$ with size α at $\mu = 0$ and power $1-\beta$ at $\mu = \delta$, as described on page 3. We allow a maximum of K analyses, with groups of n observations being taken between analyses. Each observation X_i is independent and has distribution $N(\mu, \sigma^2)$ with σ^2 known. The test is defined by choosing a value of the parameter Δ, which determines constants c_1 and c_2 as described below. Once these constants are determined, the number of observations per group, n, is given by

$$n = \left\{\frac{\sigma(c_1 + c_2)}{\delta}\right\}^2 K^{2\Delta-2}.$$

At analysis i, the total number of observations is $n_i = ni$ and we calculate $S_{n_i} = \sum_{j=1}^{n_i} X_j$. We then reject H_0 if $S_{n_i} > b_i$, accept H_0 if $S_{n_i} < a_i$ and continue to analysis $i+1$ if $S_{n_i} \in [a_i, b_i]$, where a_i and b_i are given by

$$\begin{aligned} a_i &= \delta n_i - c_2 i^{\Delta-1/2}\sqrt{n_i\sigma^2} \qquad (2.1) \\ \text{and} \qquad b_i &= c_1 i^{\Delta-1/2}\sqrt{n_i\sigma^2}, \end{aligned}$$

where $a_K = b_K$ to ensure that H_0 will be either accepted or rejected at analysis K,

should the trial proceed that far.

The constants c_1 and c_2 are determined by the choice of Δ, and can be found from tables in the paper by Pampallona & Tsiatis (1994). Alternatively, they can be found by fixing Δ and then carrying out a numerical search over possible values of c_1 and c_2 to find the values of these constants which give a test with the desired error probabilities α and β.

If the trial is to be designed for a sequence of non-equal group sizes, one possible adaptation is to replace the term $i^{\Delta-1/2}$ in the definitions of a_i and b_i (equations (2.1)) by $(Kn_i/n_K)^{\Delta-1/2}$. This reduces to equations (2.1) in the case of equal group sizes.

2.3 The error spending method

The error spending (or alpha spending) approach was proposed by Lan & DeMets (1983). Unlike earlier schemes (Pocock, 1977; O'Brien & Fleming, 1979), the error spending method does not require the number of analyses and the number of observations taken at each analysis to be specified in advance. In our notation, the maximum number of analysis K, and the sample sizes at which these analyses will be performed, $\{n_1, \ldots, n_K\}$ need not be specified as part of the trial design. An intermediate design, by Slud & Wei (1982), required K to be fixed but did not require specification of the $\{n_1, \ldots, n_K\}$.

The error spending approach does not require a fixed maximum number of analyses K and allows the analyses to be carried out after an unspecified number of observations. This may be appropriate if we wish to analyse the data every 6 months and are unsure as to the anticipated rate of accrual of information, or in a survival study where the information level is proportional to the number of observed events, which may be unpredictable. The number of observations at analysis i is still denoted as n_i, and the maximum number of analyses as K, but we note that the values K and $\{n_1, \ldots, n_K\}$

are not fixed in advance.

Instead of fixing K and $\{n_1, \ldots, n_K\}$, we specify a maximum number of observations n_{max}. We also select an error-spending function α^*, which has $\alpha^*(0) = 0$ and $\alpha^*(t) = \alpha$ for any $t \geq 1$, where α is the type I error of the test. Then the cumulative type I error at analysis i is given by $\alpha^*(n_i/n_{max})$. Thus, once the target sample size n_{max} is exceeded, the total type I error of the test is α. While it is not a formal requirement that $\alpha^*(t)$ be continuous, discontinuous error spending functions can be problematical (Li & Geller, 1991; Proschan, Follman & Waclawiw, 1992).

In §2.3.1, we describe how a two-sided test is determined by α^* and n_{max}, as was initially proposed by Lan & DeMets. In §2.3.2 we go on to describe the alterations needed to design and implement a one-sided error spending test, while §2.3.3 defines several possible choices for α^*, including two rich families of error spending functions.

2.3.1 Two-sided error spending tests

The error spending function α^* and n_{max} are used to determine the stopping boundary as follows. Consider testing the null hypothesis H_0:$\mu = 0$ against the two-sided alternative H_2:$\mu \neq 0$ with desired type I error α and error spending function α^*. Observations $X_1, X_2, \ldots$ can be taken, and $X_i \sim N(\mu, \sigma^2)$ independently of each other. At the first analysis, n_1 observations will have been taken yielding the summary statistic value $S_{n_1} = \sum_{j=1}^{n_1} X_j$. Then H_0 will be rejected if $S_{n_1} < -c_1$ or $S_{n_1} > c_1$, where c_1 is defined by $\mathbb{P}_0\{|S_{n_1}| > c_1\} = \alpha_1$, where $\alpha_1 = \alpha^*(n_1/n_{max})$. Thus, α_1 type I error is "spent" at analysis one.

In general, if the trial reaches analysis i, H_0 is rejected if $|S_{n_i}| > c_i$, where

$$\mathbb{P}_0\left\{|S_{n_1}| < c_1 \cap \ldots \cap |S_{n_{i-1}}| < c_{i-1} \cap |S_{n_i}| > c_i\right\} = \alpha^*\left(\frac{n_i}{n_{max}}\right) - \alpha^*\left(\frac{n_{i-1}}{n_{max}}\right),$$

spending $\alpha^*(n_i/n_{max}) - \alpha^*(n_{i-1}/n_{max})$ type I error at analysis i. This continues until either $|S_{n_i}| > c_i$ for some i, in which case H_0 is rejected, or $n_i \geq n_{max}$. If the maximum number of observations has been taken and the trial terminated at analysis k with $|S_{n_k}| < c_k$, H_0 is accepted.

Using the error spending approach for a two-sided test ensures that the size of the test will be α. The power of the test will be determined by the sequence of analysis times $\{n_1, \ldots, n_{max}\}$. If a particular power is required, the analysis times must be chosen appropriately, necessitating a numerical search which utilises the fact that increasing n_{max} will increase the power of the test, while reducing n_{max} will reduce the power of the test if the ratios $\{n_i/n_j; i < j \leq K\}$ are kept constant. Further details add to the complexity of this search; these are discussed, in the one-sided context, in §2.3.2.

2.3.2 One-sided error spending tests

We shall first consider the design of a one-sided error spending test, then discuss the practical implementation of the test design.

Design

To adapt the error spending approach for a one-sided test, two error spending functions must be specified. We shall use this method to test the null hypothesis H_0:$\mu \leq 0$ against the alternative H_1:$\mu > 0$ with type I error α and type II error β at $\mu = 0$ and $\mu = \delta$ respectively. This necessitates the specification of two error spending functions, $\alpha^*(t)$ and $\beta^*(t)$, which may be identical or different.

We propose a planned maximum number of analyses K and sequence of information levels $t_0 = 0 < t_1 < \ldots < t_K = 1$. Then at analysis i, the sample size is given by $n_i = t_i n_{max}$. Note that the trial can then be carried out in a more flexible manner and the planned analysis schedule deviated from, but this will perturb the power of

the test, as discussed below. At analysis i, the trial will be stopped and H_0 rejected if $S_{n_i} > b_i$; similarly the trial will be stopped and H_0 accepted if $S_{n_i} < a_i$, where a_i and b_i are given by

$$\mathbb{P}_0\{S_{n_1} \in (a_1, b_1) \cap .:. \cap S_{n_{i-1}} \in (a_{i-1}, b_{i-1}) \cap S_{n_i} > b_i\} = \alpha^*(t_i) - \alpha^*(t_{i-1}) \quad (2.2)$$

and

$$\mathbb{P}_\delta\{S_{n_1} \in (a_1, b_1) \cap \ldots \cap S_{n_{i-1}} \in (a_{i-1}, b_{i-1}) \cap S_{n_i} < a_i\} = \beta^*(t_i) - \beta^*(t_{i-1}). \quad (2.3)$$

Having selected our error spending functions and specified a planned schedule of analyses, we find n_{max} to ensure that the boundaries converge at the end of the trial if the planned schedule of analyses is adhered to. If n_{max} is larger than necessary, the boundaries will cross, while if n_{max} is too small the boundaries will not converge. As a starting point, we note that $n_{max} > n_{fix}$, thus the upper and lower boundaries of a design with $n_{max} = n_{fix}$ will not converge at analysis K. Successively larger values of n_{max} are considered until one is found where the boundaries cross, giving $a_K > b_K$. These maximum sample sizes form a bracket for a simple numerical search for the value of n_{max} such that $a_K = b_K$.

Implementation

Despite having proposed a schedule of analyses in the design of a trial using the error spending approach, we are not constrained to follow this schedule precisely. This is the advantage of the error spending approach, but it does lead to some small complications in the implementation of the trial and to perturbations in the achieved power.

In practice, the boundaries may fail to converge precisely at the first analysis where $n_i \geq n_{max}$, either crossing before this point or failing to converge at all. In either case,

the usual response is to act to preserve the type I error of the test. We can do this by finding the upper boundary point b_i as normal from equation (2.2), but instead of using equation (2.3) to determine a_i, we set $a_i = b_i$. This will result in the achieved power being greater than nominal if the boundaries had crossed and a_i is reduced and less than nominal if the boundaries had failed to converge.

One further possible complication exists. Consider the situation where, after analysis i, $\bar{\alpha}_i = \alpha - \alpha^*(n_i/n_{max})$ type I error remains to be spent, but the continuation region (a_i, b_i) is such that

$$\mathbb{P}_0\{S_{n_1} \in (a_1, b_1) \cap \ldots \cap S_{n_i} \in (a_i, b_i)\} < \bar{\alpha}_i.$$

In such a case it will not be possible to find b_{i+1} such that

$$\mathbb{P}_0\{S_{n_1} \in (a_1, b_1) \cap \ldots \cap S_{n_i} \in (a_i, b_i) \cap S_{n_{i+1}} > b_{i+1}\} = \bar{\alpha}_i.$$

An analogous situation can occur if there is less continuation probability than the unspent type II error probability. Whether this will occur at analysis i can be checked once the number of observations seen at analysis i is known but before the data is analysed. We calculate the lower and upper limits of the continuation region, a_i and b_i, as usual and if $a_i \geq b_i$ we can terminate the trial by setting $a_i = b_i$.

2.3.3 Possible error spending functions

Several possible error spending functions have been proposed in the literature. A one-sided error spending function was proposed by Jennison (1987), but this function is complicated and requires a three-dimensional search to find parameter values. Eales & Jennison (1992) have proposed using a one-sided error spending function formed by finding the error spent at each analysis under optimal schemes and interpolating these values. This is discussed in §3.4.3. The functions discussed below were all proposed in

the two-sided context, but can be used in the one-sided case.

A family of error spending functions $\alpha^*_\rho(t,\rho) = \alpha t^\rho, \rho > 0$ incorporates several functions considered in the literature, including $\rho = 1$ (Lan & DeMets, 1983) and $\rho = 1.5$ and 2 (Kim & DeMets, 1987). We shall refer to these error spending functions as the ρ-family of error spending functions.

We also consider a second family of error-spending functions which we refer to as the γ-family, proposed by Hwang, Shih & DeCani (1990). These error spending functions are defined by

$$\alpha^*_\gamma(\gamma,t) = \begin{cases} \alpha\left(1-e^{-\gamma t}\right)/\left(1-e^{-\gamma}\right) & \gamma \neq 0 \\ \alpha t & \gamma = 0. \end{cases}$$

Hwang, Shih & DeCani note that the choice of γ will depend on the situation, suggesting values between -5 and 4. Chang, Hwang & Shih (1998) use this error spending function with $\gamma = 4$ in a one-sided example and refer to its use in the Scandinavian Simvastatin Survival Study (1993), a clinical trial involving the treatment of coronary artery disease.

In the two-sided context, error spending functions exist which emulate the behaviour of the Pocock (1977) and O'Brien & Fleming (1979) repeated significance test schemes. Of these, the O'Brien & Fleming function is based on the probability of a Brownian motion crossing a horizontal boundary, which is the continuous time analogue of the O'Brien & Fleming scheme, while the Pocock function is empirically chosen to give behaviour close to that of Pocock's scheme for up to 8 interim analyses. These functions were first used by Lan & DeMets (1983) and are defined here as $\alpha^*_P(t)$ and $\alpha^*_O(t)$ respectively

$$\alpha^*_P(t) = \alpha\log\{1+(e-1)t\} \qquad \alpha^*_O(t) = 2\{1-\Phi(Z_{\alpha/2}t^{-1/2})\}.$$

Proschan (1999) investigates theoretical properties of these error spending functions when used to define two-sided tests. He finds that $\alpha^*_O(t)$ gives better results in terms

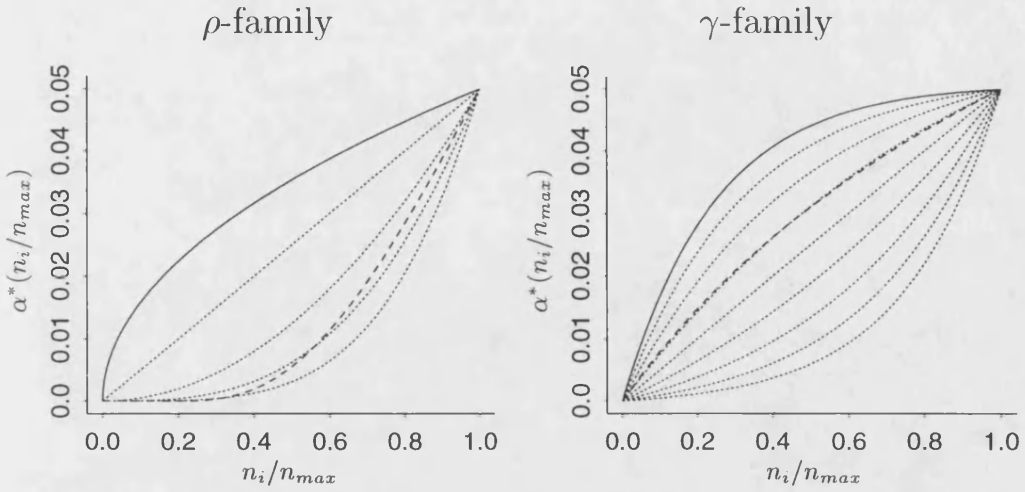


Figure 2-1: *Sample error spending functions from the γ- and ρ-families of such functions. The pseudo-Pocock and psuedo-O'Brien & Fleming error spending functions are also shown.*

of the resulting boundaries being more robust to the number of looks taken at the data as the trial progresses. Jennison & Turnbull (1990) note that the ρ-family produces error spending functions close to $\alpha_P^*(t)$ and $\alpha_O^*(t)$ with $\rho = 0.8$ and 3 respectively.

These error spending functions are shown in figure 2-1 for comparison. The ρ family error spending functions shown are for $\rho = 0.5$, 1, 2, 3 and 4 and the γ-family error spending functions are for $\gamma = -4, -3, \ldots, 4$. The solid lines on these graphs are the error spending functions with $\rho = 0.5$ and $\gamma = -4$, with other values of ρ and γ being represented by dotted lines. The dashed lines on these graphs are the pseudo-O'Brien & Fleming error spending function (on the ρ-family graph) and the pseudo-Pocock error spending function (on the γ-family graph). These functions are very close to members of the ρ- and γ-families, as can be seen.

Chapter 3

Optimal symmetric group sequential tests

Finding optimal group sequential tests is a highly computationally intensive task. Simply searching over the space of possible boundaries is not feasible for medium to large K (Pocock, 1982; Jennison, 1987). In §3.1, we describe a method due to Eales & Jennison (1992), following work by Lai (1973), which uses an artificially constructed Bayesian decision problem for which the optimal Bayes strategy corresponds to the optimal group sequential test we wish to find. In §3.1.1 the frequentist problem is defined and notation introduced, while §3.1.2 introduces the Bayesian problem identified with this frequentist question and §3.1.3 gives details of the backward induction algorithm used to solve the Bayesian problem. The performance of the optimal group sequential tests thus found is discussed in §3.2 and finally in §3.3 – §3.5 the optimal tests are used to assess the performance of the existing group sequential test designs introduced in chapter 2.

3.1 Finding optimal group sequential tests via a Bayes decision problem

3.1.1 Statistical formulation and optimality criteria

Assume that the available data are a realisation of the variates $X_1, X_2, \ldots$. Each observation is independent and identically $N(\mu, \sigma^2)$ distributed, with σ^2 assumed to be known. Up to K analyses may be carried out; the trial must terminate at the Kth analysis if not before then. Define n_i to be the cumulative number of observations at the ith analysis. For now, we consider only the case where K and $\{n_1, \ldots, n_K\}$ are known in advance as a part of the trial design. The requirement of a fixed sequence of group sizes will be relaxed in chapter 5 and in chapter 6 we will consider tests when the number of analyses is not fixed in advance. We also define the summary statistic $S_n = \sum_{j=1}^{n} X_j$ which has the distribution $S_n \sim N(n\mu, n\sigma^2)$. Finally, define the random variable N to be the actual number of observations taken up to the point when the trial terminates; then N can take values in the set $\{n_1, \ldots, n_K\}$, with $\mathbb{P}_\mu\{N = n_i\} = \mathbb{P}_\mu\{\text{trial terminates at analysis } i\}$. Suppose we wish to test $H_0{:}\mu \leq 0$ against $H_1{:}\mu > 0$, with both type I and type II error probability equal to α. We fix the type I error at $\mu = 0$ and the type II error is fixed at $\mu = \delta$, where δ is a medically significant value of treatment efficacy chosen as part of the design of the clinical trial.

We consider a subset of all possible group sequential tests, those which are symmetric about $\mu = \delta/2$. The test is defined by a set of values $\{c_1, \ldots, n_K\}$ with $c_i \geq 0$ for $i = 1, \ldots, K$. These values form the group sequential test boundary and determine the action taken at each analysis as follows.

If	$S_{n_i} < \frac{n_i\delta}{2} - c_i$	STOP, accept H_0,
if	$\frac{n_i\delta}{2} - c_i < S_{n_i} < \frac{n_i\delta}{2} + c_i$	continue to next analysis,
and if	$\frac{n_i\delta}{2} + c_i < S_{n_i}$	STOP, reject H_0.

We set $c_K = 0$ to ensure termination of the trial at the Kth analysis, if the trial has not already terminated at an earlier time.

As stated above, this group sequential test is symmetric about the intermediate value of $\mu = \delta/2$, a value of μ "midway" between H_0 and H_1. This symmetry can be seen more clearly in the notation used by Eales & Jennison (1992). In their paper, $H_1{:}\mu = -\delta$ was tested against $H_2{:}\mu = +\delta$, in which case the symmetry is about $\mu = 0$. It is easy to switch between these two situations by a simple transformation, as is noted in the paper by Eales & Jennison.

Following Jennison (1987) and Eales & Jennison (1992), we define objective functions F_1–F_5 in our notation as follows.

$$\begin{aligned}
F_1 &= \mathbb{E}_{\delta/2}\{N\} \\
F_2 &= \mathbb{E}_0\{N\} = \mathbb{E}_\delta\{N\} \\
F_3 &= \mathbb{E}_{3\delta/2}\{N\} \\
F_4 &= \frac{1}{5}\sum_{i=2}^{6} \mathbb{E}_{i\delta/4}\{N\} \\
F_5 &= \int \mathbb{E}_\mu\{N\}\frac{2}{\delta}\phi\left(\frac{2\mu-\delta}{\delta}\right)\mathrm{d}\mu
\end{aligned}$$

Given $\delta, \sigma, \alpha, K$ and $\{n_1, \ldots, n_K\}$ a group sequential test can be found which minimises any one of these objective functions. The choice of objective function to minimise will depend on the plausible treatment difference and on the treatment difference for which early stopping is most desirable. Objective function F_1 is the expected sample size if there is a small difference in favour of the new treatment; this is the value about which the test is symmetric. The two expectations given in the definition of F_2 are the expected sample size for the values of μ at which the type I and type II error rate are set. These two expectations are equal due to the symmetric nature of the tests

considered. Objective function F_3 represents a situation where the treatment effect is in fact greater than hypothesised. This objective function is also the expected sample size if $\mu = -\delta/2$, that is if the new treatment is actually inferior to the control, due to the symmetry of the problem. Objective function F_4 is averaged over μ values ranging from a small positive treatment difference to greater than predicted difference in favour of the new treatment. Objective function F_5 is the expected sample size averaged over a range of μ values weighted by a normal distribution with mean $\delta/2$ and standard deviation $\delta/2$.

Minimising certain objective functions was considered for small K (up to 10) by Jennison (1987), who searched over the space of possible $\{c_1, \ldots, c_K\}$ to find optimal tests. The following method, by Eales & Jennison (1992) and following work by Lai (1973), is more stable, faster in execution and able to deal with K up to 200. In their paper, Eales & Jennison worked with equal group sizes $n_i = in_K/K$ for $i = 1, \ldots, K$. However, the generalisation to fixed but non-equal group sizes is immediate and obvious; the following description is presented with this generalisation in place and in terms of our problem of testing $H_0{:}\mu \leq 0$ against $H_1{:}\mu > 0$.

3.1.2 The Bayes decision problem

Say we wish to minimise $F_2 = \mathbb{E}_0\{N\}$. Eales & Jennison (1992) did this by means of an artificially constructed Bayes decision problem. Define a prior on μ as $\pi(\mu) = 1/2$ for $\mu = 0$ or δ, with zero prior probability elsewhere and let the cost of sampling one observation be $c(\mu) = 1$ for $\mu = 0$ or δ and $c(\mu) = 0$ otherwise. The decision problem has two possible decisions; D_0 : '$\mu = 0$' and D_δ : '$\mu = \delta$'. The losses associated with these decisions are $L_1(D_0, \delta) = L_1(D_\delta, 0) = d$ and $L_1(D, \mu) = 0$ otherwise.

The expected cost of this decision problem is

$$\begin{aligned}
\mathbb{E}\{\text{cost}\} &= \mathbb{E}\{\text{cost of sampling}\} + \mathbb{E}\{\text{cost of decision}\} \\
&= c(0)\mathbb{E}_0\{N\}\pi(0) + c(\delta)\mathbb{E}_\delta\{N\}\pi(\delta) + \\
&\quad d\mathbb{P}_\delta\{D_0\}\pi(\delta) + d\mathbb{P}_0\{D_\delta\}\pi(0) \\
&= \frac{1}{2}\left[\mathbb{E}_0\{N\} + \mathbb{E}_\delta\{N\} + d\mathbb{P}_\delta\{D_0\} + d\mathbb{P}_0\{D_\delta\}\right] \\
&= F_2 + \frac{1}{2}d\left[\mathbb{P}_\delta\{D_0\} + \mathbb{P}_0\{D_\delta\}\right].
\end{aligned}$$

If the solution to the Bayes decision problem satisfies $\mathbb{P}_\delta\{D_0\} = \mathbb{P}_0\{D_\delta\} = \alpha$, then this solution also gives the optimal group sequential test desired. Any other decision rule for selecting between D_0 and D_δ with the stated probabilities of an incorrect choice must have a greater expected cost as this is the Bayes rule. Thus, these other rules must have greater expected sample sizes under $\mu = 0$ or δ. As each possible decision rule corresponds to a unique group sequential test, the Bayes rule uniquely identifies the group sequential test with the desired error rate α minimising F_2.

To identify the Bayes rule associated with the group sequential tests with equal error rates α, the appropriate Bayes rule is found for a fixed error cost d. The error rates for the corresponding group sequential test can be quickly calculated by means of numerical integration of the joint distribution of $\{S_{n_1}, \ldots, S_{n_K}\}$. Efficient computation for this integration is discussed by Jennison (1994). A given cost d thus leads to an achieved error probability and a search can be carried out over d to achieve the desired error probability α. For any specific value of d, the relevant test can be quickly calculated via a backward induction algorithm described in §3.1.3.

Other objective functions can be minimised by defining the loss function L_1 as for F_2 and altering the prior and cost of sampling functions $\pi(\mu)$ and $c(\mu)$ as follows. For F_1, we take $\pi(\mu) = 1/3$ for $\mu = 0$, $\delta/2$ or δ and zero otherwise, with $c(\delta/2) = 1$ and

$c(\mu) = 0$ for $\mu \neq \delta/2$. For F_3, we take $\pi(\mu) = 1/4$ for $\mu = -\delta/2$, 0, δ, and $3\delta/2$ and zero otherwise, with $c(\mu) = 1$ for $\mu = -\delta/2$ and $\mu = 3\delta/2$, with $c(\mu) = 0$ otherwise. For F_4, $\pi(\mu) = 1/10$ at $\mu = i\delta/4$ for $i = -2, \ldots, 6$ except for $\pi(\delta/2) = 1/5$. We also take $c(\mu) = 1$ for all μ for F_4. Objective function F_5 requires a more complex formulation. We take a prior on μ with probability mass 1/3 at $\mu = 0$ and δ and with a density $2\delta^{-1}\phi\{(2\mu - \delta)/\delta\}/3$, where ϕ is the density function of a standard normal variable, on $\mathbb{R}$, with $c(\mu) = 1$ for all $\mu \in \mathbb{R}\backslash\{0, \delta\}$ and $c(\mu) = 0$ for $\mu = 0$ or δ.

3.1.3 The backwards induction algorithm

Let the posterior probability that $\mu = \mu_o$ at analysis i with the observed data summarised as S_{n_i} be written as $p^{(i)}(\mu_0|S_{n_i})$. If we are to stop and make a decision at analysis i, we shall decide in favour of whichever alternative has the higher posterior probability. Hence we shall decide D_0 if $p^{(i)}(0|S_{n_i}) > p^{(i)}(\delta|S_{n_i})$ and D_δ if $p^{(i)}(\delta|S_{n_i}) > p^{(i)}(0|S_{n_i})$. Thus, the expected loss of taking this decision when the observed cumulative sum of the first n_i data values is s_{n_i} is given by

$$\gamma^{(i)}(s_{n_i}) = d\min\left\{p^{(i)}(0|s_{n_i}), p^{(i)}(\delta|s_{n_i})\right\} \qquad \text{for } i = 1, \ldots, K.$$

Let $f^{(i+1)}(s_{n_{i+1}}|s_{n_i})$ and $F^{(i+1)}(s_{n_{i+1}}|s_{n_i})$ be the probability density function and cumulative distribution function respectively of $S_{n_{i+1}}$, the sum of observations at analysis $i+1$, given an observed value s_{n_i} of S_{n_i}. We note that for $i = 2, \ldots, K$.

$$\begin{aligned}\int dF^{(i+1)}(s_{n_{i+1}}|s_{n_i}) &= \int f^{(i+1)}(s_{n_{i+1}}|s_{n_i})ds_{n_{i+1}} \\ &= \int \left\{\sum_{\mu\in\mathcal{M}} p^{(i)}(\mu|s_{n_i}) f_\mu^{(i+1)}(s_{n_{i+1}}|s_{n_i})\right\} ds_{n_{i+1}},\end{aligned}$$

where $\mathcal{M}$ is the set of values given positive prior probability by the prior π on μ; this set alters depending on the objective function specified; for example, if we wish to

minimise F_2, $\mathcal{M} = \{0, \delta\}$.

At analysis $K-1$, with an observed value $s_{n_{K-1}}$ of $S_{n_{K-1}}$, the expected cost of continuing to the next analysis and proceeding optimally there is

$$\begin{aligned}\beta^{(K-1)}(s_{n_{K-1}}) &= \mathbb{E}\{\text{cost of taking sample } K\} + \\ &\qquad \mathbb{E}\{\text{cost of acting optimally at analysis } K\} \\ &= (n_K - n_{K-1}) \sum_{\mu \in \mathcal{M}} c(\mu) p^{(K-1)}(\mu | s_{n_{K-1}}) + \\ &\qquad \int_{-\infty}^{\infty} \gamma^{(K)}(s_{n_K}) dF^{(K)}(s_{n_K} | s_{n_{K-1}}).\end{aligned} \tag{3.1}$$

Then for $i = 1, \ldots, K-2$, the expected cost of continuing to the next analysis and proceeding optimally there is given by

$$\begin{aligned}\beta^{(i)}(s_{n_i}) &= \mathbb{E}\{\text{cost of taking sample } i+1\} + \\ &\qquad \mathbb{E}\{\text{cost of acting optimally at analysis } i+1\} \\ &= (n_{i+1} - n_i) \sum_{\mu \in \mathcal{M}} c(\mu) p^{(i)}(\mu | s_{n_i}) + \\ &\qquad \int \min\{\beta^{(i+1)}(s_{n_{i+1}}), \gamma^{(i+1)}(s_{n_{i+1}})\} dF^{(i+1)}(s_{n_{i+1}} | s_{n_i}).\end{aligned} \tag{3.2}$$

We start by setting $c_K = 0$; this is done to ensure termination at the final analysis if the trial has not already concluded by then. The symmetry of the problem ensures that the final boundary point is half way between the values 0 and $n_K\delta$, the means of the distributions of S_{n_K} under D_0 and D_δ respectively. Then we search for the positive value c_{K-1} where $\gamma^{(K-1)}(n_{K-1}\delta/2 + c_{K-1}) = \beta^{(K-1)}(n_{K-1}\delta/2 + c_{K-1})$. If $\gamma^{(K-1)}(n_{K-1}\delta/2) < \beta^{(K-1)}(n_{K-1}\delta/2)$ then there is no point where the expected costs of stopping and continuing are equal. In this situation the expected cost of stopping and making a decision is less than that of continuing optimally, even if the data support each hypothesis equally strongly. In this case we set $c_{K-1} = 0$ and proceed

to analysis c_{K-2}. In order to evaluate the integral in equation (3.2) numerically when $i = K-2$, the function $\beta^{(K-1)}(S_{n_{K-1}})$ is evaluated on a grid of points in the region $[n_{K-1}\delta/2 - c_{K-1}, n_{K-1}\delta/2 + c_{K-1}]$. This enables us to evaluate $\beta^{K-2}(S_{n_{K-2}})$ and find c_{K-2}. This iterative procedure is carried out for each analysis in turn, proceeding backwards to finding c_1. This is summarised below in algorithm 1.

The search for the boundary point c_i at analysis i rely upon monotonicity of $\gamma^{(i)}(s_{n_i}) - \beta^{(i)}(s_{n_i})$. This monotonicity has been proven by Lai (1973) for the situation when we are minimising F_1 and by Brown, Cohen & Strawderman (1981) when we are minimising F_2, while we assume this property for other objective functions. Eales (1991) reports that numerical checks have supported this assumption and, like Eales, we have found no counterexamples where this monotonicity does not hold. In the two-sided setting, Chang (1996) has shown that this monotonicity holds under certain conditions. Consider testing $H_0{:}\mu = 0$ against $H_2{:}\mu \neq 0$ with type II error fixed at $\mu = \pm\delta_2$, then Chang shows that the monotonicity we require holds when the expected sample size is being minimised given any value of $\mu \in [-\delta_2, \delta_2]$.

ALGORITHM 1: SYMMETRIC FIXED GROUPS ALGORITHM

- Set $c_K = 0$.
- Find c_{K-1} by searching for $s_{n_{K-1}}$ such that $\gamma^{(K-1)}(s_{n_{K-1}}) = \beta^{(K-1)}(s_{n_{K-1}})$.
- Evaluate $\beta^{(K-1)}(s_{n_{K-1}})$ on a grid of values of $s_{n_{K-1}}$ from $n_{K-1}\delta/2 - c_{K-1}$ to $n_{K-1}\delta/2 + c_{K-1}$.
- For $i = K-2, \ldots, 2$
  - Find c_i by searching for s_{n_i} such that $\gamma^{(i)}(s_{n_i}) = \beta^{(i)}(s_{n_i})$.
  - Evaluate $\beta^{(i)}(s_{n_i})$ on a grid of values of s_{n_i} from $n_i\delta/2 - c_i$ to $n_i\delta/2 + c_i$.
- Find c_1 by searching for s_{n_1} such that $\gamma^{(1)}(n_1) = \beta^{(1)}(s_{n_1})$.

3.2 Performance of the optimal symmetric tests

We now consider the reduction in expected sample size which can be achieved by the optimal tests discussed in §3.1. In §3.2.1 we discuss the expected sample sizes of the optimal tests found via the Bayesian decision theory problem described in §3.1. We then discuss the performance of each optimal test with respect to the objective functions for which the tests have not been optimised in §3.2.2. It is desirable that the optimal tests be robust as an important factor in choosing a test may be the performance of the design with respect to unexpected treatment efficacy.

3.2.1 Optimal reduction in expected sample size

Here, we present some results indicating the reductions in expected sample size achieved by the optimal group sequential tests calculated by the method set out in this chapter. The results in table 3.1 of optimal values of F_1, F_2, F_3, and F_5 have already been presented by Eales & Jennison (1992) and in more detail in Eales (1991), but are given here for comparing with the efficiency gains possible by other group sequential schemes. The optimal values of F_4 presented in table 3.2 are not printed in these references, but correspond to and expand upon figures in Jennison (1987). The results for objective functions F_1, F_2, F_3, and F_5 are discussed in more detail by Eales (1991).

Tables 3.1 and 3.2 show the objective function values attained by the optimal tests. The tests are specified by K and t, where K is the maximum number of analyses and these analyses occur after equally sized groups of observations are taken. The maximum sample size n_K is determined by t as follows. Define n_{fix} to be the required number of observations for a fixed sample size test of H_0:$\mu \leq 0$ against H_1:$\mu > 0$ with type I error probability α at $\mu = 0$ and power $1 - \alpha$ fixed at $\mu = \delta$. If observations are independent

and identically $N(\mu, \sigma^2)$ distributed, with σ^2 known, n_{fix} is given by

$$n_{fix} = \left\{ \Phi^{-1}(1-\alpha)\frac{\sigma}{\delta} \right\}^2 .$$

where Φ is the cumulative density function of the standard normal distribution.

We then determine the maximum sample size of the group sequential test to be $n_K = t n_{fix}$. Note that these values n_K and n_{fix} are not restricted to the integers as they have a wider interpretation than as number of observations. In more complicated cases which have the same asymptotic joint distribution as the normal data case, n is more correctly interpreted as information level than sample size and takes positive real values, as discussed on page 4. Although we have presented results for t up to 1.6, we are primarily interested in $t \leq 1.2$ as higher values of t result in tests where the maximum sample size is considerably higher than the fixed sample size. If we are interested in early stopping of the trial, it is preferable not to allow the possibility of the trial continuing to the point where, for example, over 150% of the fixed sample size of observations have been taken.

The tabulated values are $100 \times F_r / n_{fix}$, for $r = 1, \ldots, 5$, for tests with the the specified maximum sample size and K equally spaced analyses. All the results given in this chapter are for tests with type I and type II error probabilities equal to 0.05. So, for example, from table 3.1 a test with 5 interim analyses and $n_K = 1.5 n_{fix}$ which is optimal for F_1 (the expected sample size when $\mu = \delta/2$) has F_1 equal to 79% of the fixed sample size.

Several general patterns apply to the results for all the objective functions considered. If t is fixed, increasing K reduces the optimum value of the objective function. McPherson (1982) reported results where increasing K increased the expected sample size, but this was for a class of schemes where altering K caused a change in n_K. When the maximum sample size is kept constant, in all the examples we have seen, increasing K

	K				
t	2	5	10	15	20
1.01	93.4	87.9	85.7	84.9	84.5
1.05	88.9	82.5	80.0	79.1	76.1
1.10	87.3	80.2	77.5	76.5	76.1
1.15	**87.0**	79.1	76.2	75.3	74.8
1.20	87.2	78.6	75.6	74.5	74.0
1.30	88.3	**78.4**	75.0	73.9	73.3
1.40	90.0	78.6	**74.9**	**73.7**	**73.1**
1.50	91.7	79.0	75.0	**73.7**	**73.1**
1.60	93.6	79.5	75.3	73.8	**73.1**

F_1

	K				
t	2	5	10	15	20
1.01	80.9	72.2	69.1	68.1	67.6
1.05	74.5	65.2	62.1	61.0	60.5
1.10	72.8	62.2	59.0	57.9	57.4
1.15	**72.7**	60.7	57.4	56.2	55.7
1.20	73.2	59.8	56.3	55.2	54.6
1.30	75.3	59.0	55.2	53.9	53.3
1.40	78.1	**58.7**	54.6	53.3	52.6
1.50	81.2	58.8	54.4	53.0	52.2
1.60	84.7	59.2	**54.3**	**52.8**	**52.0**

F_2

	K				
t	2	5	10	15	20
1.01	59.7	48.6	45.1	44.0	43.4
1.05	**57.0**	41.7	38.4	37.3	36.7
1.10	57.9	39.1	35.6	34.4	33.8
1.15	59.5	37.8	34.0	32.9	32.3
1.20	61.5	37.2	33.1	31.8	31.2
1.30	65.9	**36.8**	31.9	30.6	29.9
1.40	70.5	37.1	31.3	29.8	29.2
1.50	75.3	37.7	31.0	29.4	28.7
1.60	80.2	38.6	**30.8**	**29.1**	**28.3**

F_3

	K				
t	2	5	10	15	20
1.01	83.3	75.9	73.1	72.2	71.7
1.05	78.2	69.6	66.7	65.7	65.2
1.10	**76.8**	67.0	63.9	62.9	62.4
1.15	**76.8**	65.7	62.5	61.4	60.9
1.20	77.4	65.0	61.7	60.6	60.0
1.30	79.4	**64.5**	60.9	59.7	59.0
1.40	81.9	64.6	**60.6**	59.3	58.6
1.50	84.7	64.9	**60.6**	**59.2**	**58.5**
1.60	87.7	65.4	60.7	**59.2**	**58.5**

F_5

Table 3.1: *Optimal values of the specified objective functions, given as percentages of the sample size required for the equivalent non-sequential test, for sequential designs with K equally spaced analyses. Maximum sample size is t times the fixed sample size, type I error is 0.05 and power is 0.95. The bold figures are the minimum values over t for each fixed K.*

	K									
t	2	3	4	5	10	15	20	30	50	100
1.01	78.7	74.4	72.2	70.7	67.8	66.8	66.4	65.9	65.5	65.4
1.05	73.8	68.2	65.8	64.3	61.3	60.3	59.8	59.3	59.0	58.7
1.10	**72.9**	65.9	63.2	61.7	58.5	57.5	57.0	56.5	56.1	55.8
1.15	73.2	65.1	62.0	60.4	57.1	56.0	55.5	54.9	54.5	54.2
1.20	74.0	**64.9**	61.5	59.7	56.3	55.1	54.6	54.0	53.6	53.3
1.30	76.5	65.4	**61.3**	**59.2**	55.4	54.2	53.6	53.0	52.5	52.2
1.40	79.4	66.5	61.7	**59.2**	**55.1**	53.8	53.1	52.5	52.0	51.6
1.50	82.6	68.0	62.4	59.6	**55.1**	**53.7**	**52.9**	**52.2**	51.7	51.3
1.60	86.0	69.7	63.3	60.2	55.2	**53.7**	**52.9**	**52.2**	**51.6**	**51.1**

Table 3.2: *Optimal values of F_4, given as percentages of the sample size required, for the equivalent non-sequential test for sequential designs with K equally spaced analyses. Maximum sample size is t times the fixed sample size, type I error is 0.05 and power is 0.95. The bold figures are the minimum values over t for each fixed K.*

reduces the optimum objective function values. However, the improvement attained by adding an extra interim analysis is modest for over roughly 5 analyses. In practice, the increased logistical problems caused by frequent interim analyses will preclude the use of a very large number of analyses; it is for precisely this reason that group sequential schemes have been developed as an alternative to fully sequential methods. While the improvement in efficiency made possible by increasing the maximum number of analyses is only an observed pattern for general increases in K, multiplying K by some integer constant will, at worst, result in the same optimal values of the objective functions. To see this, consider two group sequential tests which are equivalent in terms of all parameters other than the number of analyses. Assume that one test has a maximum of K_1 analyses, while the second test has a maximum of kK_1 analyses, where k is a positive integer, with both tests having the interim analyses equally spaced throughout the trial. The set of all possible tests with kK_1 analyses includes the set of all possible tests with a maximum of K_1 looks at the data. In particular, the optimal test with a maximum of K_1 analyses can be considered as a test with kK_1 analyses when the boundaries at certain analysis times are infinite. Thus, the test with the larger number

of analyses must have optimum objective function values no worse than the test with a smaller number of analyses.

For any fixed K and any specific objective function, there is a value of t which will minimise the objective function over the t values considered. In general, the objective function at first decreases as t is increased until this minimising value is found, and as t increases further the objective function will increase again. However, most of these minimising values of t are larger (in many cases much larger) than we would wish to consider for practicable test designs, as discussed above. Thus, selecting a test design can require a trade-off between minimising the expected sample size and keeping the maximum sample size within reasonable limits. However, there is a large reduction in expected sample sizes if a maximum sample size only a little greater than n_{fix} is allowed. The further reduction in expected sample size achieved by having a larger maximum sample size is in most cases relatively modest compared to the benefit of allowing a maximum sample size a little larger than n_{fix}.

The lowest achievable objective function values are for F_3, with the expected sample size in this case being below half the fixed sample size with only 5 analyses for as little as a 1% increase in the maximum sample size over n_{fix}. This is attributable to the fact that F_3 refers to the situation where there is a large positive treatment effect or a small negative effect. These situations are the most extreme that we consider, and it is not surprising that under these circumstances it is possible to stop the trial very early. Objective functions F_2, F_4 and F_5 have reductions of similar magnitude to each other, all being more modest than those achieved by F_3, while the smallest reductions are for F_1. Minimising F_1 is the Keifer-Weiss problem (Weiss, 1962) of minimising the expected sample size under the worst situation with respect to the true treatment efficacy, as $\mu = \delta/2$ is the value of μ where the minimum expected sample size is maximised. That is, for any $\mu \neq \delta/2$, $\mathbb{E}_{\mu}\{N\} < \mathbb{E}_{\delta/2}\{N\}$.

3.2.2 Performance of optimal tests with respect to other objective functions

This section explores the behaviour of the optimal symmetric tests with respect to the objective functions for which they have not been optimised. Throughout this thesis, we shall refer to a test optimised for objective function F_r as OPT_r.

Figure 3-1 shows how well the various optimal tests do in terms of achieved values of the objective functions for which the tests are not optimised. For each graph, the horizontal axis gives the maximum sample size of the tests as a percentage of the fixed sample size n_{fix} while the vertical axis gives the achieved value of the objective function in question, again as a percentage of the fixed sample size. The separate graphs are for F_1 to F_5, as indicated by the labels on the vertical axes. Values for tests with a maximum sample size larger than 140% of n_{fix} have been excluded, for the reasons discussed in the previous section. All the tests used to produce this figure had $K = 10$ equally spaced analyses; similar patterns of results were seen for $K = 5, 15$ and 20.

It can clearly be seen that the OPT_3 tests perform worst for the other objective functions, having the highest achieved values for all other objective functions. The graph of achieved F_3 values also shows that it is for this objective function that the tests optimised for other objective functions are furthest from being optimal. However, even the worst of them (OPT_1) has achieved values of F_3 which are below 40% of n_{fix}.

The OPT_2, OPT_4 and OPT_5 tests all achieve very nearly optimal results with respect to objective functions F_2, F_4 and F_5, while OPT_1 performs worse than this trio for all objective functions other than F_1.

Overall, it seems that OPT_1 should only be considered as a candidate design if placing all emphasis on a small positive treatment difference is deemed sensible. Similarly, OPT_3 tests are only sensible if the most important consideration is the behaviour of the test when $\mu = -\delta/2$ or $\mu = 3\delta/2$.

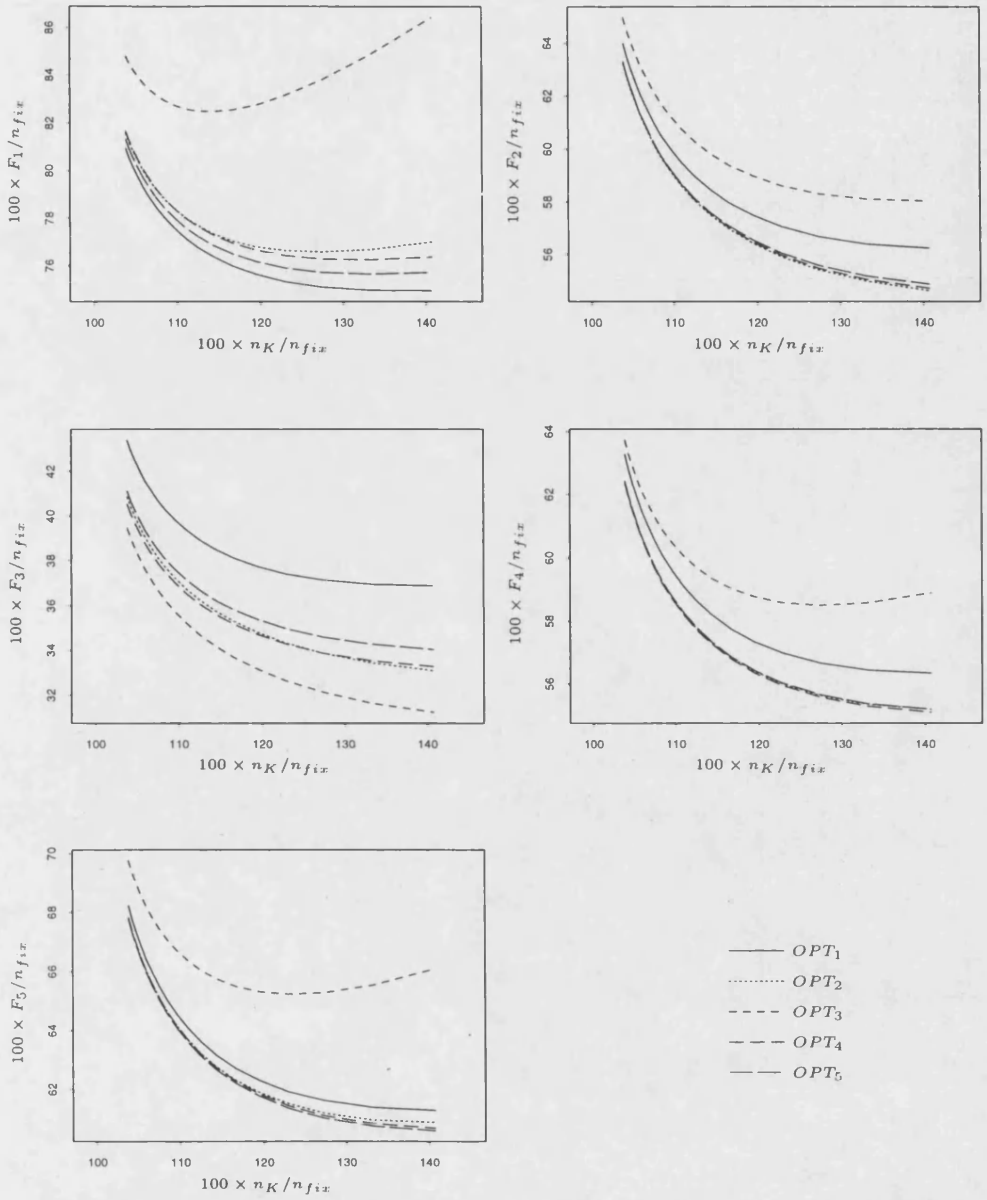


Figure 3-1: *Achieved values of F_1 to F_5 for tests optimising these objective functions. All tests have type I error 0.05 and power 0.95 with 10 equally spaced analyses. Maximum sample size and objective function values are given as percentages of the fixed sample size n_{fix}.*

Conversely, the OPT_2, OPT_4 and OPT_5 tests perform well over the range of objective functions. All of these require similar computing power to find the optimal tests, but the evaluation of F_5 requires considerably more computing power to evaluate, due to the extra numerical integration required for dealing with a prior which consists of both point masses and a continuous density. Since the performance of all the tests we have considered is similar for both F_4 and F_5, OPT_4 seems a good choice for a test when good expected sample size properties are required across a range of objective functions, without requiring the extra computation involved by considering F_5.

It would be simple to alter the prior and cost of sampling functions used in the Bayes problem which was defined in §3.1.2 to produce tests which optimise other objective functions. However, before such tests were used, it would be wise to consider the behaviour of such tests with respect to other objective functions as has been done in this section.

3.3 Performance of the Δ-family

In section §2.2, we introduced the Δ-family of group sequential tests. In this section, we use the optimal tests examined in the preceding section to assess the performance of the Δ-family tests with respect to their expected sample sizes.

Table 3.3 shows the maximum sample size and achieved values of objective functions F_1 – F_5 for Δ-family tests of H_0:$\mu \leq 0$ against H_1:$\mu > 0$ with type I error probability 0.05 at $\mu = 0$ and power 0.95 fixed at $\mu = \delta$. The results are for K = 2, 5, 10, 15 and 20 equally spaced analyses and for $\Delta = -0.5, -0.3, -0.1, 0.1$ and 0.3. Recall that the maximum sample sizes of the tests are determined by the choice of Δ; these values are given as percentages of the fixed sample size n_{fix}. For each value of Δ we compute the optimal group sequential tests which minimise F_1 to F_5 amongst all group sequential tests of the same hypotheses with the same error rates and group sizes. The tabulated

	K				
Δ	2	5	10	15	20
-0.5	100.0	101.9	103.9	104.9	105.6
-0.3	100.3	103.4	106.0	107.3	108.2
-0.1	101.5	106.2	109.6	111.4	112.4
0.1	104.8	111.8	116.7	119.1	120.5
0.3	111.9	125.1	133.2	137.1	139.5

Maximum sample sizes.

	K				
Δ	2	5	10	15	20
-0.5	100.0	102.2	104.2	105.3	105.9
-0.3	100.0	102.1	103.8	104.7	105.2
-0.1	100.0	101.7	103.1	103.7	104.1
0.1	100.0	101.0	101.8	102.1	102.4
0.3	100.0	100.2	100.5	100.6	100.6

Relative performances for F_1.

	K				
Δ	2	5	10	15	20
-0.5	100.0	109.2	113.8	115.9	117.2
-0.3	100.0	108.6	112.5	114.3	115.4
-0.1	100.0	107.3	110.4	111.9	112.7
0.1	100.0	105.4	107.4	108.4	109.0
0.3	100.0	102.7	103.6	104.0	104.2

Relative performances for F_2.

	K				
Δ	2	5	10	15	20
-0.5	100.0	133.1	143.1	148.1	151.1
-0.3	100.0	129.5	139.1	143.3	145.9
-0.1	100.0	126.3	133.3	136.8	138.8
0.1	100.0	120.7	125.6	127.9	129.4
0.3	100.0	110.0	115.9	117.0	117.7

Relative performances for F_3.

	K				
Δ	2	5	10	15	20
-0.5	100.0	110.3	114.4	116.3	117.5
-0.3	100.0	109.2	112.7	114.4	115.3
-0.1	100.0	107.6	110.3	111.6	112.3
0.1	100.0	105.4	106.9	107.7	108.1
0.3	100.0	102.1	102.6	102.8	102.9

Relative performances for F_4.

	K				
Δ	2	5	10	15	20
-0.5	100.0	107.2	110.7	112.3	113.3
-0.3	100.0	106.5	109.5	110.9	111.7
-0.1	100.0	105.4	107.7	108.7	109.4
0.1	100.0	103.7	105.0	105.6	106.0
0.3	100.0	101.3	101.7	101.9	101.9

Relative performances for F_5.

Table 3.3: *Tabulated values are* $100 \times n_K/n_{fix}$ *(top left) and relative performances for objective functions* F_1 – F_5 *achieved by* Δ*-family tests with type I error 0.05, power 0.95 and* K *equally spaced analyses.*

results are the values of F_1 to F_5 achieved by the Δ-family tests, given as percentages of the optimal objective function values. In the discussions in this and later sections, we refer to the percentage comparison of achieved and optimal objective function values as relative performance.

The maximum sample sizes of these tests increase as as Δ increases and as K increases, and this trend continues for tests with larger K and Δ than than those shown in table 3.3. In order to keep the maximum sample size at a reasonable level, Δ must be chosen appropriately. For small K and negative Δ, the maximum sample sizes are quite good

and are still reasonable for low K and small positive Δ.

The most striking point of the relative performances shown in table 3.3 is the fact that all tests with $K = 2$ have relative performance of 100%, i.e. they achieve the same objective function values as the optimal tests. This is because, with the final boundary point fixed by the symmetric nature of the problem, the error rates of the test can only be altered by the placing of the boundary points at the first analysis. This means that there is only one symmetric two-analysis test of our hypothesis which has the desired error rates, and hence all methods find this single two-analysis test.

For all the tests shown, the relative performance of the Δ-family tests improves with respect to all the objective functions as Δ increases. This means that improvement in relative performance is gained at the cost of increasing the maximum sample size of the test. The best relative performances are for F_1, with good to fair performance for all values of K and Δ. Considering the relatively poor performance of the OPT_1 tests with respect to F_3 in §3.2.2, it is therefore not surprising to see that the Δ-family tests perform especially poorly with respect to F_3. The relative performances for F_2, F_4 and F_5 are reasonable to poor, with the best being for F_5. With $\Delta = 0.3$, the Δ-family tests are very nearly optimal for F_1 and near optimal for F_2, F_4 and F_5, but this performance comes at the cost of a large maximum sample size.

The advantage of the Δ-family tests over the optimal tests we have considered lies in their ease of calculation. Once the parameter Δ is chosen and the constants c_1 and c_2 determined, either by numerical searches or by consulting tables such as those given by Pampallona & Tsiatis (1994), the remaining calculations are simple. However, these relative performances make it clear that this simplicity of design comes at the cost of poor performance for F_3 and somewhat inadequate performance with respect to F_2, F_4 and F_5. The good relative performance for F_1 makes the Δ-family a sensible choice with regards to the expected sample size in this situation.

3.4 Performance of the error spending method

In this section, we use the optimal tests discussed in this chapter to assess the performance of error spending tests using the ρ-family and γ-family error spending functions described in §2.3.3. All the error spending tests referred to in this section have been designed for a sequence of K equally spaced analyses, as described in §2.3.2. However, they do of course retain the flexibility of the error spending method and can be applied to other sequences of group sizes, although this will perturb the achieved power and efficiency of these tests.

3.4.1 The γ-family of error spending tests

Table 3.4 shows maximum sample size and relative performance results for the error spending tests using the γ-family error spending functions for $\gamma = -4, -3, \ldots, 1$. The maximum sample size figures are given as percentages of the fixed sample size and the relative performance figures are the objective function values achieved by the γ-family tests percentages of the optimal objective function values for matching tests, as defined in the previous section.

The maximum sample size figures are reasonably low for negative values of γ, but increase with increasing γ and are higher for larger K. Maximum sample size figures for values of γ larger than those tabulated rapidly become unreasonably large.

Relative performance values for F_1 are very good and are slightly better for lower γ, corresponding to smaller maximum sample sizes. This is in contrast to all the other objective functions studied, where the relative performances are better for higher γ, corresponding to larger maximum sample sizes. The relative performance of the γ-family tests is good for F_5, although not as good as for F_1, and good to fair for F_2 and F_4. The figures for F_3 show reasonable relative performance.

γ	K = 2	5	10	15	20
-4	101.5	103.9	105.4	106.1	106.4
-3	102.8	106.5	108.4	109.2	109.6
-2	105.0	110.4	113.0	113.9	114.4
-1	108.6	116.4	119.7	120.9	121.5
0	113.9	125.0	129.3	130.8	131.6
1	121.1	136.5	142.2	144.2	145.2

Maximum sample sizes.

γ	K = 2	5	10	15	20
-4	100.0	100.4	100.9	101.2	101.3
-3	100.0	100.4	100.9	101.2	101.3
-2	100.0	100.4	101.0	101.3	101.4
-1	100.0	100.5	101.2	101.5	101.7
0	100.0	100.6	101.6	102.0	102.2
1	100.0	100.8	102.2	102.8	103.1

Relative performances for F_1.

γ	K = 2	5	10	15	20
-4	100.0	101.4	102.8	103.4	103.8
-3	100.0	101.2	102.5	103.0	103.3
-2	100.0	101.1	102.1	102.6	102.8
-1	100.0	100.9	101.7	102.2	102.4
0	100.0	100.7	101.4	101.8	102.1
1	100.0	100.5	101.0	101.5	101.8

Relative performances for F_2.

γ	K = 2	5	10	15	20
-4	100.0	106.7	109.4	110.8	111.5
-3	100.0	106.5	108.6	109.7	110.4
-2	100.0	106.1	107.8	108.7	109.3
-1	100.0	105.7	107.2	107.9	108.4
0	100.0	105.1	106.7	107.3	107.7
1	100.0	104.4	106.4	106.8	107.2

Relative performances for F_3.

γ	K = 2	5	10	15	20
-4	100.0	101.2	102.3	102.9	103.2
-3	100.0	101.0	101.9	102.4	102.6
-2	100.0	100.8	101.5	101.9	102.1
-1	100.0	100.6	101.1	101.4	101.7
0	100.0	100.3	100.7	101.1	101.3
1	100.0	100.1	100.4	100.8	101.1

Relative performances for F_4.

γ	K = 2	5	10	15	20
-4	100.0	100.7	101.7	102.2	102.4
-3	100.0	100.6	101.4	101.8	102.0
-2	100.0	100.5	101.2	101.5	101.7
-1	100.0	100.3	100.9	101.3	101.5
0	100.0	100.2	100.7	101.1	101.3
1	100.0	100.1	100.6	101.1	101.4

Relative performances for F_5.

Table 3.4: *Tabulated values are* $100 \times n_K/n_{fix}$ *(top left) and relative performances for objective functions* F_1 – F_5 *achieved by error spending tests using the* γ*-family of error spending function with type I error 0.05, power 0.95 and* K *equally spaced analyses.*

The combination of lower maximum sample size and better relative performance make it clear that the lowest values of γ are best if the minimisation of F_1 is the primary goal in a sequential design using the γ-family. At these lower values of γ, the relative performance for F_2, F_4 and F_5 is also good. However, if any other objective function than F_1 is of primary importance, it is necessary to strike a balance between maximum sample size and relative performance. For tests with a maximum of 5 or 10 analyses, choosing $\gamma = -2$ or -3 will give a test design with good to reasonable maximum sample size and expected samples sizes within about 4% of optimal for objective functions F_1, F_2, F_4 and F_5.

3.4.2 The ρ-family of error spending tests

Maximum sample size and relative performance results for tests using the ρ-family error spending function with $\rho = 1, 2, 3$, and 4 are shown in table 3.5. The maximum sample size increases as ρ decreases and as K increases. Once more, relative performance figures are good for F_1, with the best values for $\rho = 2$. All other objective functions show better relative performance for low values of ρ than for high, corresponding to better relative performance when the maximum sample size is larger. Objective function F_5 shows good relative performance, while results for F_2 and F_4 are good to fair. Once more, the results for F_3 are less impressive, although better than for the Δ-family and γ-family tests.

For tests with a maximum of 5 or 10 analyses, a test with maximum sample size within about 10% of the fixed samples size can be gained by choosing $\rho = 2$ or 3. This choice gives good relative performance with respect to objective functions F_1, F_2, F_4 and F_5, achieving objective function values within 4% of optimal in these cases.

	K				
ρ	2	5	10	15	20
1.0	113.9	125.0	129.3	130.8	131.6
2.0	104.5	110.1	112.7	113.7	114.2
3.0	101.6	105.0	106.9	107.6	108.0
4.0	100.6	102.6	104.1	104.7	105.0

Maximum sample sizes.

	K				
ρ	2	5	10	15	20
1.0	100.0	100.6	101.6	102.0	102.2
2.0	100.0	100.2	100.5	100.6	100.7
3.0	100.0	100.4	100.9	101.0	101.2
4.0	100.0	100.6	101.2	101.5	101.7

Relative performances for F_1.

	K				
ρ	2	5	10	15	20
1.0	100.0	100.7	101.4	101.8	102.1
2.0	100.0	102.0	102.6	102.9	103.1
3.0	100.0	102.9	104.1	104.6	104.8
4.0	100.0	103.4	105.2	105.9	106.3

Relative performances for F_2.

	K				
ρ	2	5	10	15	20
1.0	100.0	105.1	106.7	107.3	107.7
2.0	100.0	110.3	111.7	112.3	112.7
3.0	100.0	113.1	115.3	116.4	116.9
4.0	100.0	114.4	118.0	119.5	120.4

Relative performances for F_3.

	K				
ρ	2	5	10	15	20
1.0	100.0	100.3	100.7	101.1	101.3
2.0	100.0	101.9	102.2	102.4	102.5
3.0	100.0	103.0	103.8	104.2	104.4
4.0	100.0	103.6	105.1	105.7	106.0

Relative performances for F_4.

	K				
ρ	2	5	10	15	20
1.0	100.0	100.2	100.7	101.1	101.3
2.0	100.0	101.1	101.4	101.6	101.7
3.0	100.0	101.9	102.6	103.0	103.1
4.0	100.0	102.4	103.6	104.1	104.4

Relative performances for F_5.

Table 3.5: *Tabulated values are* $100 \times n_K/n_{fix}$ *(top left) and relative performances for objective functions* $F_1 - F_5$ *achieved by error spending tests using the* ρ*-family of error spending function with type I error 0.05, power 0.95 and* K *equally spaced analyses.*

3.4.3 Error spending functions defined from optimal tests

A number of authors have proposed forming an error spending function by evaluating the error spent at each analysis by some existing repeated significance test scheme and interpolating a function through these points, either as a linear interpolation or in some more complicated fashion. If the accrual of information was in accordance with that anticipated by the repeated significance test scheme in question, using the error spending function thus defined would recover the original test. This would allow greater flexibility in dealing with information accrual which was not as anticipated.

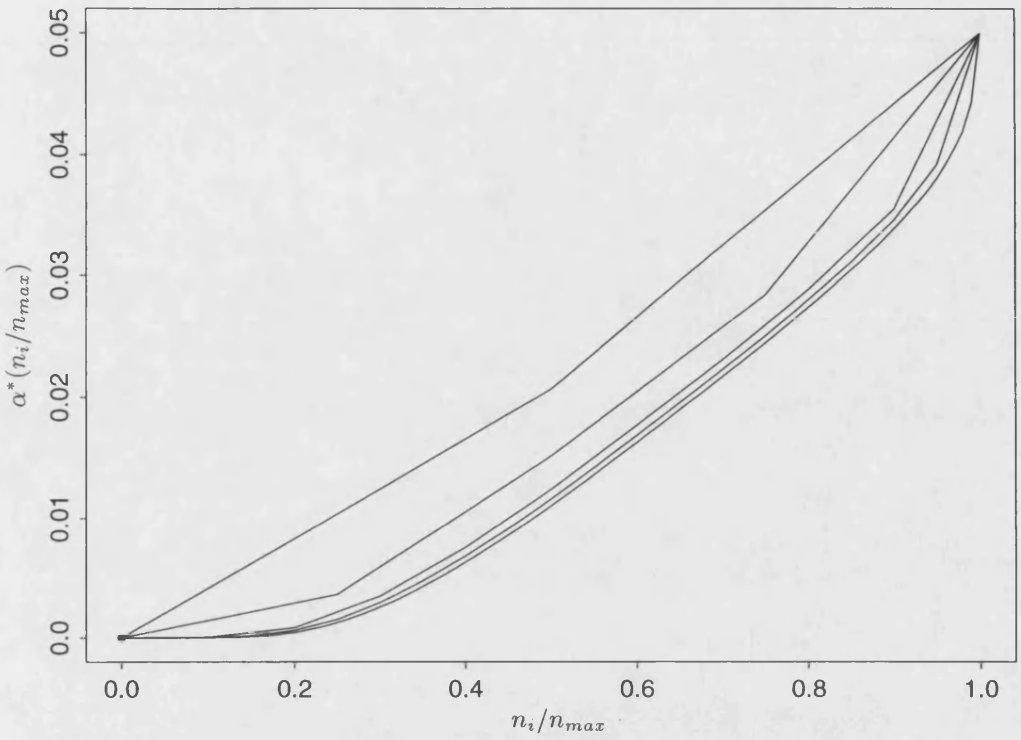


Figure 3-2: *Error spending functions defined by interpolating the error spent by optimal tests. Tests are optimised for F_3, with $n_K = 1.01 \times n_{fix}$ and for $K = 2$ (uppermost line), 4, 10, 20 and 100 (lowermost line).*

An example is given in Eales & Jennison (1992), where the error spending function is defined by finding the OPT_2 test with 100 analyses and using using linear interpolation. This error spending function is then used to carry out a test with 6 analyses. The relative performance of the resulting test for F_2 is approximately 101.4%, with n_K being approximately 110% of n_{fix}. The nearest comparable case from tables 3.4 and 3.5 show relative performance of 102.0% for $\rho = 2.0$ or 101.1% for $\gamma = -2$, although these are not directly comparable figures as Eales & Jennison's example has 6 unequal group sizes and the results from tables 3.4 and 3.5 are for 5 equally spaced analyses. However, this does illustrate that the simple parametric error spending functions are

highly efficient if the value of ρ or γ is chosen with care.

This approach assumes that there is an underlying optimal error spending function and that the implicit error spending function derived by interpolation from an optimal test will be an approximation to this. Further, by using an optimal test with a large number of analyses, it is hoped to make this estimate of the optimal error spending function as smooth as possible. Figure 3-2 shows the error spending functions which are implied by tests optimised for F_3 when $T = 1.01$ and for $K = 2$, 4, 10, 20 and 200. There are clear differences between these error spending functions for different values of K, implying that the error spending function defined by the optimal test with large K may not be the most suitable for trials where only a small number of analyses will be carried out, as was observed by Eales & Jennison.

3.5 Comparing the Δ-family, γ-family and ρ-family tests

Figure 3-3 compares the performance of the Δ-family, γ-family and ρ-family tests directly. The plots are of the maximum sample size against the achieved objective function values, both given as percentages of the fixed sample size n_{fix}. Optimal values for each objective function are included for reference. The dotted line represents Δ-family tests for which $\Delta = -0.5, -0.45, \ldots, 0.3$, with tests which had a maximum sample size of over 135% of n_{fix} excluded. Similarly, γ- and ρ-family tests in this figure are for $\gamma = -5.0, -4.5, \ldots, 0.5$ and $\rho = 0.9, 1.0, \ldots, 4.0$

Looking at the results for the maximum sample size being no more than 120% of n_{fix}, it is clear that the Δ-family tests are the furthest from optimal, while the two families of error spending tests have similar performance, with the the γ-family being slightly superior For objective functions F_2 to F_5, although the ρ-family tests are slightly superior to the γ-family tests for F_1.

In practice, the relative simplicity of the Δ-family tests is a large advantage. Once the

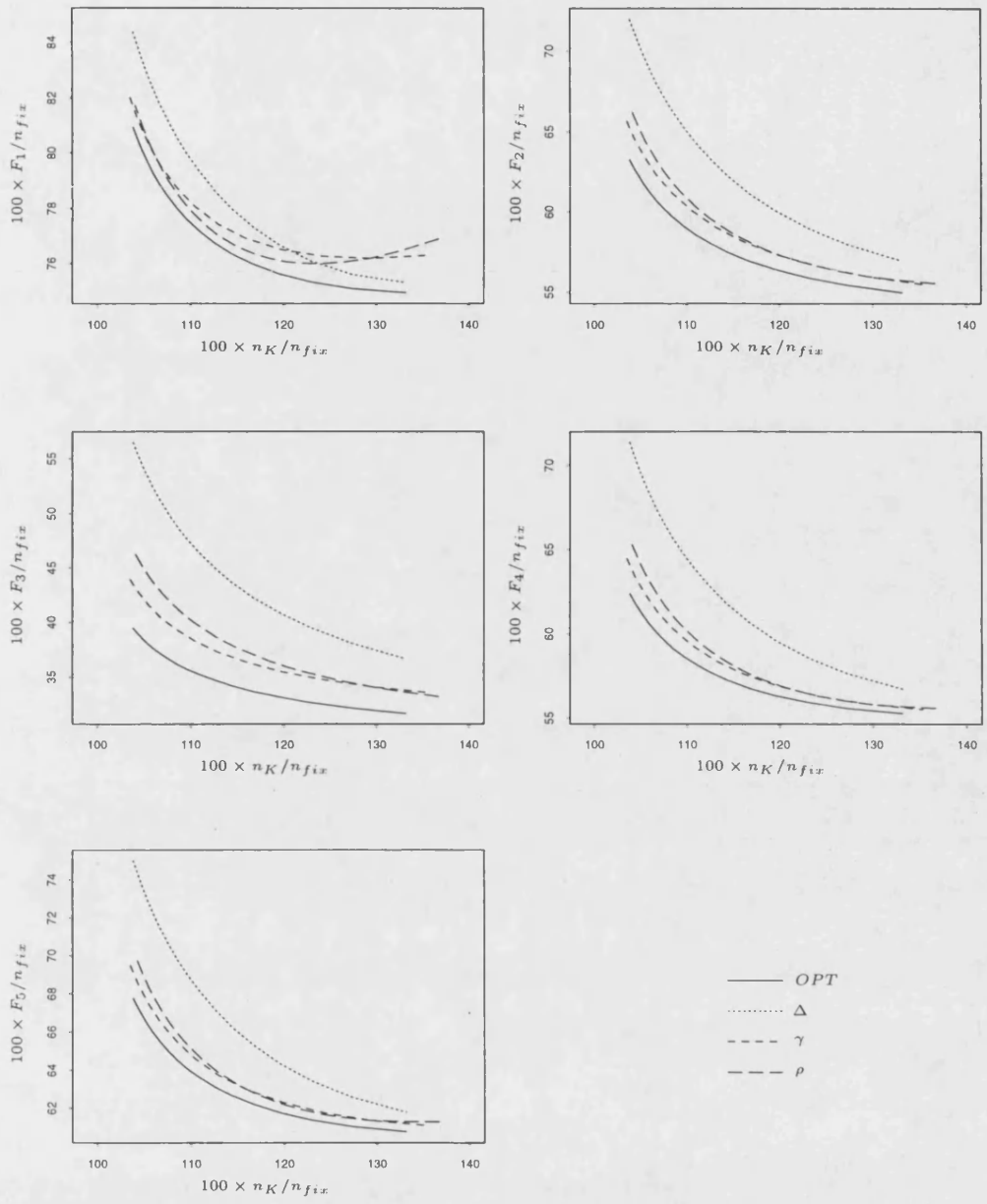


Figure 3-3: *Achieved values of F_1 to F_5 for Δ-family, γ-family and ρ-family tests. All tests have type I error 0.05 and power 0.95, with 10 equally spaced analyses. Optimal objective function values are included for comparison.*

constants c_1 and c_2 are found from tables in Pampallona & Tsiatis (1994) or Jennison & Turnbull (2000; chapter 4), the calculations involved are simple. In contrast, an error spending test involves lengthy computation and is not available in many commonly used statistical software packages. However, this advantage is overshadowed by the flexibility of the error spending method and the fact that the error spending tests have a clear performance advantage. Error spending tests using $\gamma = -3$ or -2, or using $\rho = 2$ or 3 have both good maximum sample size and expected sample size properties, as we have seen. These tests, therefore, are a good choice if the optimal group sequential tests are inappropriate due to the difficulty of specifying the schedule of analyses in advance.

Chapter 4

Optimal asymmetric group sequential tests

Without further modification, the method used in chapter 3 only finds optimal group sequential tests for symmetric problems, that is tests with equal type I and type II error probabilities. Also, the set of tests searched over to find the optimal test only includes test boundaries which are symmetric about the mean of the hypothesised differences in efficacy between the experimental and control treatments. The symmetric nature of the problem with equal error rates ensures that the optimal test is in this subset of all possible group sequential tests. In this chapter, we extend this method to find optimal group sequential tests which have unequal error rates, necessitating a search over the larger set of test boundaries which are not symmetric about the mean of the hypothesised differences in efficacy.

4.1 The induced asymmetry and some consequences

The statistical formulation of the problem given on page 23 is restated here for convenience. Recall that up to K groups of observations are taken with cumulative sample sizes $n_1, n_2, \ldots, n_K$. Each observation is independent and identically distributed with a $N(\mu, \sigma^2)$ distribution, where σ^2 is known. We wish to test $H_0{:}\mu \leq 0$ against $H_1{:}\mu > 0$ with type I error α at $\mu = 0$ and type II error β at $\mu = \delta$. We define N to be the number of observations taken by the termination of the trial, and we also define the summary statistic $S_n = \sum_{i=1}^{n} X_i$ for $n = n_1, \ldots, n_K$.

In the previous chapter, we dealt with this problem when the error rates are equal, that is when $\alpha = \beta$. However, in practice this is often not the case. For example, if the trial is examining the efficacy of a treatment for a rare medical condition or seeking a small improvement in efficacy, the required sample size will be very large and in such situations a lower power is often considered an acceptable price to pay for a more reasonable sample size. Thus, we now consider tests where the error rates are unequal.

The symmetry between the error rates in the previous chapter implied an attitude that the null and alternative hypotheses are of equal importance. This attitude is reflected in some of the objective functions defined on page 24. For example, F_2 and F_3 are defined as being

$$\begin{aligned} F_2 &= \mathbb{E}_0\{N\} = \mathbb{E}_\delta\{N\} \\ F_3 &= \mathbb{E}_{-\delta/2}\{N\} = \mathbb{E}_{3\delta/2}\{N\}. \end{aligned}$$

The symmetry of the problem ensures that the expected sample size under $\mu = 0$ equals that when $\mu = \delta$; similarly the expected sample sizes when $\mu = -\delta/2$ and $\mu = 3\delta/2$ are equal. However, these expectations are only equal if the test has $\alpha = \beta$ and is symmetric about the mean of the hypothesised differences in efficacy. In the case of asymmetric tests, we may wish to place different emphasis on the expected sample size

when $\mu = 0$ to that when $\mu = \delta$. Thus, we consider several new objective functions.

$$F_{21} = \frac{1}{2}\left(\mathbb{E}_0\{N\} + \mathbb{E}_\delta\{N\}\right) \qquad F_{22} = \mathbb{E}_0\{N\} \qquad F_{23} = \mathbb{E}_\delta\{N\}$$
$$F_{31} = \frac{1}{2}\left(\mathbb{E}_{-\delta/2}\{N\} + \mathbb{E}_{3\delta/2}\{N\}\right) \qquad F_{22} = \mathbb{E}_{-\delta/2}\{N\} \qquad F_{23} = \mathbb{E}_{3\delta/2}\{N\}$$

These objective functions replace F_2 and F_3. Note that F_{21}, F_{22}, and F_{23} reduce to F_2 in the symmetric case and in this situation, F_{31}, F_{32}, and F_{33} reduce to F_3. Of these new objective functions, F_{21} and F_{31} are the closest to the spirit of the original F_2 and F_3 proposed by Jennison (1987). The definition of F_4 on page 24 is re-written as

$$F_4 = \frac{1}{5}\mathbb{E}_{\delta/2}\{N\} + \frac{1}{10}\sum_{i=-2, i\neq 2}^{6} \mathbb{E}_{\frac{i}{4}\delta}\{N\}.$$

This is the mean expected sample size over values of μ from $-\delta/2$ to $3\delta/2$, with double weight placed on $\mu = \delta/2$. In the case of symmetric group sequential tests, this is equal to the formula given on page 24. We retain the existing definitions of F_1 and F_5;

$$F_1 = \mathbb{E}_{\delta/2}\{N\} \qquad F_5 = \int \mathbb{E}_\mu\{N\}\frac{2}{\delta}\phi\left(\frac{2\mu - \delta}{\delta}\right)\mathrm{d}\mu.$$

Allowing unequal error rates breaks the symmetry of the test boundary about the mean of the hypothesised differences in efficacy, $\delta/2$, where the continuation region at analysis i was $[n_i\delta/2 \pm c_i]$, for some $c_i \geq 0$ (page 24). Instead, the action taken at analysis i, for $i = 1, \ldots, K$ is determined by values a_i and b_i, where $a_i \leq b_i$, as follows.

If	$S_{n_i} < a_i$	STOP, accept H_0,
if	$a_i < S_{n_i} < b_i$	continue to next analysis,
and if	$b_i < S_{n_i}$	STOP, reject H_0.

To ensure termination of the trial at analysis K, we set $a_K = b_K$. Thus, the symmetric problem is a special case of this more general formulation, with $a_i = n_i\delta/2 - c_i$ and

$b_i = n_i\delta/2 + c_i$. As we shall see in the next section, this increases the complexity of both the Bayes decision theory problem which we use to find out optimal group sequential test and the backward induction algorithm used to solve the decision theory problem.

4.2 Adapting the Bayes decision problem method to asymmetric tests

Both the Bayes decision theory problem which has a solution corresponding to the optimal group sequential test we seek and the backward induction algorithm we use to find the solution to this problem must be modified to cope with the asymmetric case we discuss in this chapter. In this section, we consider the necessary modifications, first to the Bayes problem in §4.2.1, then to the backward induction algorithm in §4.2.2.

4.2.1 Alterations to the Bayes decision problem

Recall the Bayes decision problem defined in §3.1.2 for minimising F_2. We now consider finding the asymmetric group sequential test which minimises F_{21}. As in the symmetric case, we wish to choose between decisions D_0 : '$\mu = 0$' and D_δ : '$\mu = \delta$' with a prior on μ of $\pi(\mu) = 1/2$ for $\mu = 0$ or δ and zero otherwise. The cost of sampling is $c(\mu) = 1$ for $\mu = 0$ and δ, but zero otherwise, also as in the symmetric case. The loss function now changes from the symmetric case, and the cost of a wrong decision is now $L_2(D_0, \delta) = d_\delta, L_2(D_\delta, 0) = d_0$ and $L_2(D, \mu) = 0$ otherwise. The critical change from the symmetric case we discussed in chapter 3 is that the cost of the two possible wrong decisions are not constrained to be equal. This reflects our desire to place greater emphasis on avoiding one of the two possible wrong decisions, rather than being equally concerned over the two possible errors.

The expected cost of any decision rule for this problem is

$$\begin{aligned}
\mathbb{E}\{\text{cost}\} &= \mathbb{E}\{\text{cost of sampling}\} + \mathbb{E}\{\text{cost of decision}\} \\
&= c(0)\mathbb{E}_0\{N\}\pi(0) + c(\delta)\mathbb{E}_\delta\{N\}\pi(\delta) + \\
&\quad d_\delta \mathbb{P}_\delta\{D_0\}\pi(\delta) + d_0\mathbb{P}_0\{D_\delta\}\pi(0) \\
&= \frac{1}{2}\left(\mathbb{E}_0\{N\} + \mathbb{E}_\delta\{N\} + d_\delta\mathbb{P}_\delta\{D_0\} + d_0\mathbb{P}_0\{D_\delta\}\right) \\
&= \frac{1}{2}F_{21} + \frac{1}{2}\left(d_\delta\mathbb{P}_\delta\{D_0\} + d_0\mathbb{P}_0\{D_\delta\}\right).
\end{aligned}$$

Thus, if the Bayes rule minimising the total expected cost satisfies $\mathbb{P}_0\{D_\delta\} = \alpha$ and $\mathbb{P}_\delta\{D_0\} = \beta$, this Bayes rule minimises F_{21} amongst all decision strategies with the probabilities of making a wrong decision being α and β. Thus, this Bayes rule identifies the group sequential test with the desired error rates that minimises F_{21}. To find the Bayes rule with the desired probabilities of making a wrong decision, we now conduct a two-dimensional search over d_0 and d_δ in place of the one dimensional search over the cost parameter d required in the symmetric case.

As in the symmetric case, the optimisation of other objective functions is accomplished by altering the prior $\pi(\mu)$ and cost of sampling function $c(\mu)$, while the loss function $L_2(D, \mu)$ remains unchanged regardless of which objective function is being minimised. For F_{22} and F_{23}, $\pi(\mu)$ is the same as for F_{21}, but for F_{22} $c(0) = 1$ and $c(\mu) = 0$ for $\mu \neq 0$, while for F_{23} $c(\delta/2) = 1$ and $c(\mu) = 0$ for $\mu \neq \delta/2$. When minimising F_{31}, F_{32} and F_{33}, the prior takes value $1/4$ for $\mu = -\delta/2, 0, \delta/2$ and $3\delta/2$ while being zero for any other value of μ. For F_{31}, the only non-zero values of $c(\mu)$ are $c(-\delta/2) = c(3\delta/2) = 1$, for F_{32} $c(-\delta/2) = 1$ is the only non-zero value of $c(\mu)$ and for F_{33} $c(3\delta/2) = 1$ is the only non-zero value of $c(\mu)$. The objective functions F_1, F_4 and F_5 are unchanged and so the priors and sampling cost functions for these objective functions are as given on page 27, however we now use loss function $L_2(D, \mu)$ rather than $L_1(D, \mu)$.

4.2.2 Alterations to the backward induction algorithm

Most of the backward induction algorithm used to find the Bayes rule in the symmetric case in §3.1.3 remains unchanged. However, the breaking of the symmetry about $\mu = \delta/2$ does add some complications. The upper and lower boundary points at analysis i are no longer equidistant from $n_i\delta/2$, and indeed will both be above or below this point at some analyses.

As before, we start at analysis K. Recall that the posterior probability that $\mu = \mu_0$ given an observed value s_{n_K} of S_{n_K} is written as $p^{(K)}(\mu_0|s_{n_K})$. As we must stop at analysis K, if we have not done so already, we shall decide D_0 if the expected loss of this decision is less than the expected loss of deciding D_δ. That is, we shall decide D_0 if $d_\delta p^{(K)}(\delta|s_{n_K}) < d_0 p^{(K)}(0|s_{n_K})$ and D_δ if $d_0 p^{(K)}(0|s_{n_K}) < d_\delta p^{(K)}(\delta|s_{n_K})$. Thus, the point where we are balanced between the two choices is $s^*_{n_K}$, where $s^*_{n_K}$ is such that

$$d_\delta p^{(K)}(\delta|s^*_{n_K}) = d_0 p^{(K)}(0|s^*_{n_K}).$$

Solving this equation for $s^*_{n_K}$, we find that

$$s^*_{n_K} = \frac{\delta n_K}{2} - \frac{\sigma^2}{\delta}\log\left\{\frac{d_\delta}{d_0}\right\} \tag{4.1}$$

and we set $a_K = b_k = s^*_{n_K}$.

At analysis $K-1$, we can still calculate the expected cost of stopping and making a decision and the expected cost of continuing to the next analysis and acting optimally there, $\gamma^{(K-1)}(s_{n_{K-1}})$ and $\beta^{(K-1)}(s_{n_{K-1}})$ respectively, as described in §3.1.3. However, it is no longer true that the boundary points a_{K-1}, b_{K-1} are given by finding the single value c_{K-1} and setting the boundary points to $n_{K-1}\delta/2 \pm c_{K-1}$. Instead, we must find a_{K-1} and b_{K-1} separately.

If we stop the trial at analysis $K-1$, we shall decide decision D_0 if $d_\delta p^{(K-1)}(\delta|s_{n_{K-1}}) <$

$d_0 p^{(K-1)}(0|s_{n_{K-1}})$ and D_δ if $d_0 p^{(K-1)}(0|s_{n_{K-1}}) < d_\delta p^{(K-1)}(\delta|s_{n_{K-1}})$. Thus, we can find $s^*_{n_{K-1}}$ such that the expected costs of the two decisions are equal; this value can be found using equation (4.1) with the obvious alterations for analysis $K-1$ instead of analysis K and we can deduce that $a_{K-1} \leq s^*_{n_{K-1}} \leq b_{K-1}$. In order to find the boundary points, we first find $s^*_{n_{K-1}}$ and then check whether $\gamma^{(K-1)}(s^*_{n_{K-1}}) < \beta^{(K-1)}(s^*_{n_{K-1}})$. If this is the case, the expected cost of stopping and making a decision is less than the expected cost of continuing to the final analysis and continuing optimally there. In this situation, we shall stop if we observe $S_{n_{K-1}} = s^*_{n_{K-1}}$ and hence we shall stop for any observed value of $S_{n_{K-1}}$, so we set $a_{K-1} = b_{K-1} = s^*_{n_{K-1}}$.

If $\gamma^{(K-1)}(s^*_{n_{K-1}}) > \beta^{(K-1)}(s^*_{n_{K-1}})$, we search for two values of $S_{n_{K-1}}$ such that the expected costs of stopping and continuing optimally are equal and we set our boundary points to these values, with $a_{K-1} \leq s^*_{n)K-1} \leq b_{K-1}$ Having found these points, we then evaluate $\beta^{(K-1)}(S_{n_{K-1}})$ on a grid of values of $S_{n_{K-1}}$ from a_{K-1} to b_{K-1} in order to be able to evaluate $\beta^{(K-2)}(S_{n_{K-2}})$, as in the symmetric case. We then proceed backwards to analysis 1, finding $\{s^*_{n_i}, a_i, b_i\}$ and evaluating $\beta^{(i)}(S_{n_i})$ on a grid of values of $S_{n_i} \in [a_i, b_i]$ at analysis i for $i = K-2, K-3, \ldots, 1$. This is summarised in algorithm 2, below.

As in the symmetric case, the uniqueness of the boundary points (a_i, b_i) at each analysis i and the convergence of our searches for these points depend upon the monotonicity of the expected costs of stopping and continuing optimally, as discussed on page 29. Numerical checks have supported our assumption of monotonicity.

It is clear from comparing algorithm 2 to algorithm 1 on page 29 that the two are fundamentally similar. However, each single search in the case of the symmetric algorithm has been replaced by two searches here and the one-dimensional search over the cost parameter d in the symmetric case is replaced by a two-dimensional search over d_0 and d_δ in the asymmetric case.

Algorithm 2: Asymmetric fixed groups algorithm

ANALYSIS K

- Find $s^*_{n_K}$ such that $d_\delta p^{(K)}(\delta|s^*_{n_K}) = d_0 p^{(K)}(0|s^*_{n_K})$.
- Set $a_K = b_K = s^*_{n_K}$.

ANALYSES $K-1,\ldots,2$

- For $i = K-2,\ldots,2$:
  - Find $s^*_{n_i}$ such that $d_\delta p^{(i)}(\delta|s^*_{n_i}) = d_0 p^{(i)}(0|s^*_{n_i})$.
  - If $\gamma^{(i)}(s^*_{n_i}) > \beta^{(i)}(s^*_{n_i})$:
    - Find a_i such that $\gamma^{(i)}(a_i) = \beta^{(i)}(a_i)$, with $a_i < s^*_{n_i}$.
    - Find b_i such that $\gamma^{(i)}(b_i) = \beta^{(i)}(b_i)$, with $b_i > s^*_{n_i}$.
    - Evaluate $\beta^{(i)}(S_{n_i})$ on a grid of values of S_{n_i} from a_i to b_i.
  - Otherwise, set $a_i = b_i = s^*_{n_i}$.

ANALYSIS 1

- Find $s^*_{n_1}$ such that $d_\delta p^{(1)}(\delta|s^*_{n_1}) = d_0 p^{(1)}(0|s^*_{n_1})$.
- If $\gamma^{(1)}(s^*_{n_1}) > \beta^{(1)}(s^*_{n_1})$:
  - Find a_1 such that $\gamma^{(1)}(a_1) = \beta^{(1)}(a_1)$, with $a_1 < s^*_{n_1}$.
  - Find b_1 such that $\gamma^{(1)}(b_1) = \beta^{(1)}(b_1)$, with $b_1 > s^*_{n_1}$.
- Otherwise, set $a_1 = b_1 = s^*_{n_1}$.

This increased dimensionality creates a practical problem with the search for costs of wrong decisions. In the symmetric case, we had a one-dimensional search, and knew that we were seeking a positive real value. Thus, we could take zero as one end of a search bracket, and steadily increase the cost of a wrong decision until a cost was

found that gave a the probability of a wrong decision to be less than α. This gave us the other end of a search bracket, and a simple numerical algorithm could be used to find the cost of a wrong decision which resulted in the desired probability of a wrong decision. However, in the two-dimensional search arising from our asymmetric problem, it is not possible to bracket the point we seek, where the values of d_0 and d_δ result in probabilities of the two wrong decisions being α and β. Numerical checks are required for the convergence of these searches, and in the small number of cases where the search was not initially successful, alternative starting points were used.

4.3 Performance of the optimal asymmetric tests

In this section, we consider the performance of the optimal asymmetric tests. Firstly, in §4.3.1 we consider the minimum objective function values attained for objective functions F_1, F_{21}, F_{31}, F_4 and F_5. Objective functions F_1, F_4 and F_5 are unchanged from the symmetric case discussed in chapter 3, while F_{21} and F_{31} are the closest analogues to objective functions F_2 and F_3. In §4.3.2 we consider the tests optimised for F_{21}, F_{22}, and F_{23} and for F_{31}, F_{32}, and F_{33}. Finally, in §4.3.3 we discuss the performance of the optimised group sequential tests with respect to objective functions other than the ones for which they are optimal.

4.3.1 Optimal reduction in expected sample size

Table 4.1 shows the minimum values of objective functions F_1, F_{21}, F_{31}, F_4 and F_5 for tests of $H_0{:}\mu \leq 0$ against $H_1{:}\mu > 0$, with type I error probability $\alpha = 0.05$ at $\mu = 0$ and type II error probability $\beta = 0.10$ fixed at $\mu = \delta$. Up to K interim analyses are permitted, occurring after equally sized groups of observations. The maximum number of observations is $n_K = tn_{fix}$, where n_{fix} is the sample size required for the equivalent fixed-sample test. The tabulated minima of these objective functions are

	K				
t	2	5	10	15	20
1.01	92.9	87.4	85.2	84.4	84.0
1.05	88.1	81.7	79.2	78.3	77.9
1.10	86.4	79.2	76.5	75.6	75.1
1.15	**86.0**	78.0	75.2	74.2	73.7
1.20	86.2	77.4	74.4	73.4	72.8
1.30	87.3	**77.0**	73.7	72.5	72.0
1.40	88.9	77.1	**73.5**	**72.2**	71.6
1.50	90.8	77.4	**73.5**	**72.2**	**71.5**
1.60	92.7	77.8	73.7	72.3	71.6

F_1

	K				
t	2	5	10	15	20
1.01	82.9	74.7	71.8	70.8	70.3
1.05	76.6	67.7	64.7	63.7	63.2
1.10	74.8	64.7	61.5	60.5	50.0
1.15	**74.6**	63.1	59.8	58.7	58.2
1.20	75.0	62.1	58.7	57.6	57.0
1.30	76.9	61.2	57.5	56.3	55.7
1.40	79.4	**60.9**	56.9	55.5	54.9
1.50	82.4	61.0	56.6	55.2	54.5
1.60	85.6	61.3	**56.5**	**55.0**	**54.2**

F_{21}

	K				
t	2	5	10	15	20
1.01	63.6	53.0	49.5	48.4	47.9
1.05	**59.6**	45.7	42.4	41.3	40.7
1.10	59.8	42.8	39.3	38.2	37.6
1.15	61.1	41.3	37.6	36.5	35.9
1.20	62.8	40.5	36.5	35.3	34.7
1.30	66.8	**39.9**	35.2	33.9	33.2
1.40	71.2	**39.9**	34.4	33.0	32.3
1.50	75.8	40.4	34.0	32.4	31.7
1.60	80.5	41.1	**33.8**	**32.1**	**31.3**

F_{31}

	K				
t	2	5	10	15	20
1.01	80.6	72.8	69.9	69.0	68.5
1.05	75.3	66.2	63.2	62.2	61.8
1.10	**74.1**	63.3	60.3	59.2	58.7
1.15	74.2	61.9	58.7	57.6	57.1
1.20	74.9	61.1	57.8	56.7	56.1
1.30	77.1	60.5	56.8	55.5	54.9
1.40	79.9	**60.4**	56.3	55.0	54.4
1.50	83.0	60.7	**56.2**	54.8	54.1
1.60	86.2	61.2	**56.2**	**54.7**	**54.0**

F_4

	K				
t	2	5	10	15	20
1.01	84.6	77.3	74.6	73.7	73.3
1.05	79.2	70.9	68.0	67.1	66.6
1.10	77.7	68.1	65.1	64.1	63.6
1.15	**77.6**	66.7	63.6	62.5	62.0
1.20	78.0	65.9	62.7	61.6	61.0
1.30	79.8	65.3	61.7	60.5	59.9
1.40	82.3	**65.2**	61.3	60.0	59.4
1.50	85.0	65.5	**61.2**	**59.8**	**59.1**
1.60	87.8	65.9	61.3	**59.8**	**59.1**

F_5

Table 4.1: *Tabulated values are the optimal values of the specified objective functions, given as percentages of the sample size required for the equivalent non-sequential test. Sequential designs have K equally spaced analyses. Maximum sample size is t times the fixed sample size, type I error is 0.05 and power is 0.90. The bold figures are the minimum values over t for each fixed K.*

given as percentages of n_{fix}. For definitions of the objective functions, see page 50.

In the case of the symmetric tests considered in chapter 3, objective functions F_{21} and F_{31} are equal to F_2 and F_3 respectively. Thus, the values in table 4.1 can be directly compared with those in tables 3.1 and 3.2. The patterns shown by these optimal asymmetric tests are similar to those seen for the symmetric tests. In particular, for any fixed K the minimum over t of each objective function occurs at very similar values of t for the symmetric and asymmetric cases. These minimising values of t are in most cases larger than we would like, leading to group sequential tests with maximum sample sizes considerably larger than the fixed sample size, but the majority of the reduction in expected sample size can be gained with a relatively small increase in maximum sample size. For example, allowing a five analysis test to have a maximum sample size of 110% of n_{fix} will reduce the minimum achievable value of F_5 to 68.1% of n_{fix}, but raising the maximum sample size to 140% of n_{fix} only reduces the optimal values of F_5 by a further 2.9% of n_{fix}.

The achieved values of F_1 are slightly lower in the asymmetric case, but those of the other objective functions considered are higher for the asymmetric tests than for the symmetric tests. However, these differences are not large, and the relative pattern of the lowest objective function values being for F_{31}, with F_{21}, F_4 and F_5 achieving similar optima and F_1 having the largest values which was observed in the symmetric case is still true here. Overall, the results for these optimal asymmetric tests are very similar to those found in the symmetric case.

4.3.2 Asymmetric objective functions

In the previous section, we considered the behaviour of the symmetric objective functions. These objective functions are symmetric in that they give equal weight to the null and alternative hypotheses. Since we are now permitting asymmetric tests,

	K				
t	2	5	10	15	20
1.01	77.1	68.7	65.7	64.7	64.2
1.05	70.7	60.7	57.8	56.8	56.3
1.10	**69.7**	57.7	54.6	53.6	53.1
1.15	70.2	56.4	53.2	52.1	51.5
1.20	71.3	55.7	52.3	51.2	50.7
1.30	74.0	**55.4**	51.6	50.4	49.8
1.40	77.2	55.6	**51.4**	50.1	49.4
1.50	80.7	56.1	**51.4**	**50.0**	49.3
1.60	84.3	56.6	**51.4**	**50.0**	**49.2**

F_{22}

	K				
t	2	5	10	15	20
1.01	83.2	75.6	72.8	71.9	71.4
1.05	77.4	69.0	66.1	65.1	64.6
1.10	**76.2**	66.5	63.5	62.5	62.0
1.15	**76.2**	65.4	62.2	61.2	60.7
1.20	76.8	64.9	61.5	60.4	59.9
1.30	78.8	**64.4**	60.8	59.6	59.0
1.40	81.2	64.5	60.6	59.2	58.6
1.50	83.9	64.7	**60.5**	59.1	58.4
1.60	86.9	65.1	**60.5**	**59.0**	**58.3**

F_{23}

Table 4.2: *Optimal values of objective functions F_{22} and F_{23}, given as percentages of the sample size required for the equivalent non-sequential test. Sequential designs have K equally spaced analyses. Maximum sample size is t times the fixed sample size, type I error is 0.05 and power is 0.90. The bold figures are the minimum values over t for each fixed K.*

we will not always give equal importance to the expected sample sizes when $\mu = 0$ and when $\mu = \delta$, hence the introduction of objective functions F_{22}, F_{23}, F_{32} and F_{33}. In this section, we consider the behaviour of these asymmetric objective functions and discuss when they may be useful.

We shall first consider objective functions F_{21} to F_{23}. Recall that they are defined as follows.

$$F_{21} = \frac{1}{2}\left(\mathbb{E}_0\{N\} + \mathbb{E}_\delta\{N\}\right) \qquad F_{22} = \mathbb{E}_0\{N\} \qquad F_{23} = \mathbb{E}_\delta\{N\}.$$

Note that in the case of symmetric tests, these three objective functions are all equivalent to the objective function F_2 used in the discussion of symmetric tests. Objective function F_{21} retains the spirit of the symmetric F_2, and optimum values for F_{21} are in table 4.1, while equivalent values for F_{22} and F_{23} are in table 4.2. All tests were of $H_0{:}\mu \leq 0$ against $H_1{:}\mu > 0$ with type I error probability $\alpha = 0.05$ and

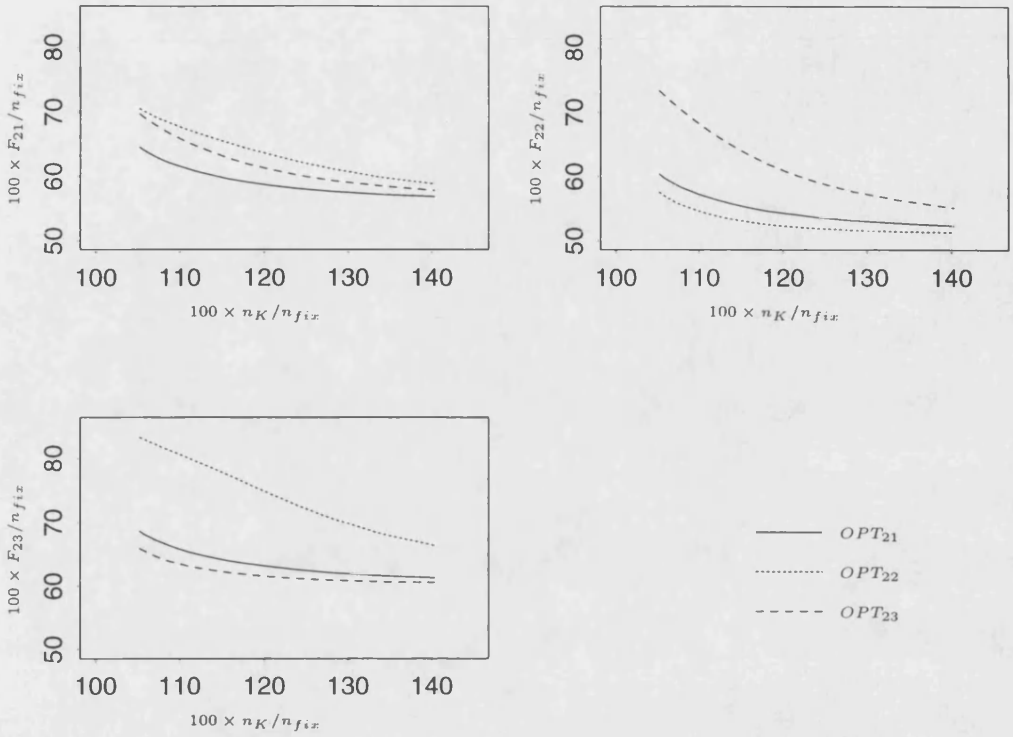


Figure 4-1: *Achieved values of F_{21}, F_{22} and F_{23} for tests optimising these objective functions. All tests have type I error 0.05 and power 0.90 with 10 equally spaced analyses. Recall that F_{21} is the average of the expected sample sizes under H_0 and H_1, while F_{22} and F_{23} are the expected sample sizes under H_0 and H_1 respectively.*

power $1 - \beta = 0.90$ fixed at $\mu = \delta$. The optimum objective function values for F_{22} and F_{23} occur at the same values of t for any fixed K, while those for F_{21} occur at slightly higher t for the same K.

The greatest achieved reduction in objective function values are for F_{22}, with F_{23} having the least reduction in achieved objective function values. The achieved efficiency gains for F_{21} are between those for F_{22} and F_{23}. The optimum values of F_{22} are lowest since they represent the expected sample size under H_0, and with $\alpha < \beta$ it is harder to reject H_0 when $\mu = 0$ than it is to accept H_0 when $\mu = \delta$. The effect of this is to move the

boundaries of the test upwards, making early stopping more likely when $\mu = 0$.

Recall that we refer to a test optimised for objective function F_r as the OPT_r test. Figure 4-1 shows the values of F_{21}, F_{22} and F_{23} achieved by the OPT_{21}, OPT_{22} and OPT_{23} tests over a range of maximum sample sizes. The horizontal axis gives the maximum sample size as a percentage of n_{fix}, the sample size required by the equivalent fixed sample test, while the vertical axis gives the achieved value of the relevant objective function, again as a percentage of n_{fix}. In each case, there was a maximum of 10 equally spaced analyses.

All three tests perform similarly with respect to F_{21}, although the difference is largest for lower maximum sample sizes. Unsurprisingly, the OPT_{22} and OPT_{23} tests do poorly with respect to objective functions F_{23} and F_{22} respectively; this is to be expected as the situation for which they are optimised differ greatly. It is, however, noteworthy that the OPT_{21} test comes close to the optimal values of both F_{22} and F_{23}. This suggests that, in general use, the OPT_{21} tests have more desirable expected sample size characteristics than the OPT_{22} and OPT_{23} tests.

There are, however, examples where the use of the OPT_{22} or OPT_{23} tests would be suitable. Figure 4-1 suggests that optimising a test for its performance under H_0 or H_1 should only be done if there is a specific reason for desiring early stopping under that hypothesis. For example, if we were examining the performance of a new treatment which has superior secondary characteristics to an established treatment, we might test H_0':$\mu \leq -\delta$ against H_1':$\mu > -\delta$, $\delta > 0$ as we would accept the new treatment so long as it was not clinically significantly worse than the existing treatment. In this situation, we would desire early stopping under H_0', as in that case the new treatment would be significantly inferior to the existing one, but we would prefer to gather more information to assess the secondary characteristics of the new treatment if we were going to recommend use of the new treatment. By suitable translating our standard problem of testing H_0:$\mu < 0$ against H_1:$\mu > 0$, we could use an OPT_{22} test in this

situation, which would stop early under H_0', when we would want early stopping, but allow a longer study of the secondary characteristics of the treatment under H_1'.

The symmetric objective function F_3 gives rise to F_{31} and the asymmetric objective functions F_{32} and F_{33} in the same way as F_2 lead to F_{21}, F_{22} and F_{23}. Examination of optimal objective function values and relative performances of F_{31}, F_{32} and F_{33} showed the same patterns as seen in table 4.2 and figure 4-1, as did other examples with different numbers of analyses for both F_{21} to F_{23} and F_{31} to F_{33}.

The optimal tests considered in this section show that tests optimised for objective function F_{21} perform extremely well with respect to their expected sample size when $\mu = 0$ or $\mu = \delta$. However, tests optimised for one of these μ values perform poorly under the other hypothesis. Hence, in the following sections we concentrate on considering tests optimised for F_{21} as in most cases these tests will be more useful than those optimised for $\mu = 0$ or $\mu = \delta$. Similarly, we consider the F_{31} tests, which are optimised for the mean of the expected sample sizes when $\mu = -\delta/2$ and $\mu = 3\delta/2$, rather than the tests optimised for $\mu = -\delta/2$ or $3\delta/2$.

4.3.3 Performance of optimal tests with respect to other objective functions

The behaviour of symmetric tests optimised for objective functions F_1 to F_5 with respect to those objective functions for which they were not optimised was explored in §3.2.2. Figure 4-2 shows similar comparisons of the performance of tests $OPT_1, OPT_{21}, OPT_{31}, OPT_4$ and OPT_5 with respect to the objective functions for which they are not optimal. All tests had 10 equally spaced analyses, with the maximum sample size as given on the horizontal axis and achieved objective function values given on the vertical axis, both as percentages of n_{fix}. Tests with other numbers of equally spaced analyses were observed to follow the same patterns.

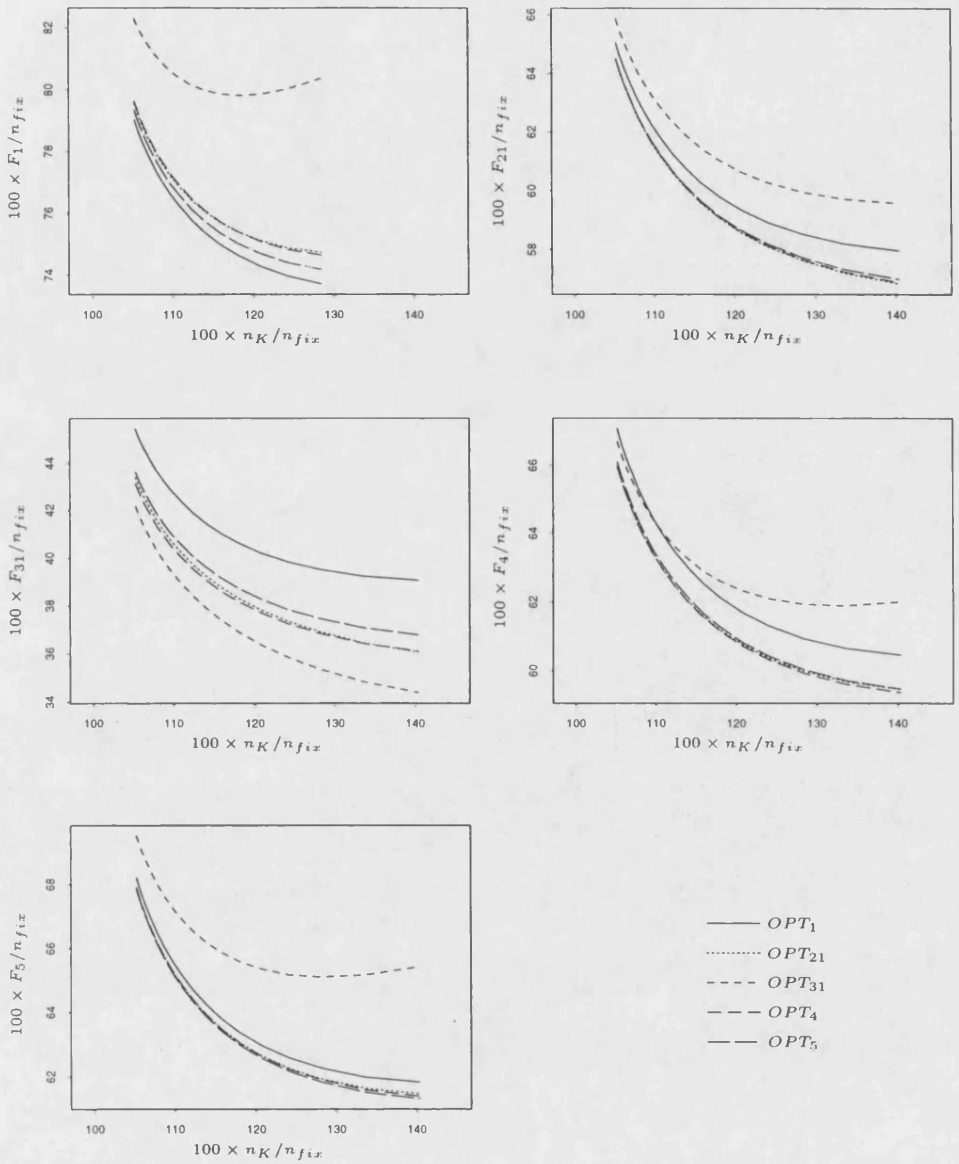


Figure 4-2: *Achieved values of F_1, F_{21}, F_{31}, F_4 and F_5 for tests optimising these objective functions. All tests have type I error 0.05 and power 0.90 with 10 equally spaced analyses. Maximum sample sizes and objective function values are given as percentages of the fixed sample size n_{fix}.*

The results in figure 4-2 are extremely similar to those seen in figure 3-1 for the symmetric case. The OPT_1 and OPT_{31} tests perform poorly for all objective functions other than those for which they are optimised and especially poorly for objective functions F_{31} and F_1 respectively. The tests optimised for F_{21}, F_4 and F_5 have similar performance in most cases, with the OPT_5 tests being slightly better with respect to F_1 and the OPT_{21} and OPT_4 tests being slightly better for F_3. Overall, the OPT_{21} and OPT_4 tests are more practically useful than the OPT_5 tests due to their lower computational burden.

4.4 Performance of the Δ-family

In this section we discuss the performance of the Δ-family of group sequential tests when the type I and II error probabilities are not equal. The performance is assessed with respect to the expected sample sizes attained, with consideration also given to the maximum sample sizes of the tests.

Table 4.3 shows the maximum sample sizes and achieved objective function values for Δ-family tests over a range of Δ and K. All tests are of H_0:$\mu \leq 0$ against H_1:$\mu > 0$, with type I error probability $\alpha = 0.05$ and power $1 - \beta = 0.90$ at $\mu 0$ and $\mu = \delta$ respectively. The maximum sample size values are given as a percentage of n_{fix}, while the achieved objective function values are given as percentages of the optimal values of the specified objective functions for tests with the same group sizes. These ratios of achieved to optimal objective function values are referred to as relative performances.

The maximum sample size of the Δ-family tests increases as Δ increases and is larger for tests with a higher maximum number of analyses. For 5 analyses, $\Delta \leq 0.1$ gives a group sequential test with maximum sample size less than 120% of n_{fix}, while for $K = 10, 15$ or 20, $\Delta \leq -0.1$ keeps the maximum sample size below this threshold.

The worst relative performance is for F_{31}, which is poor. Relative performances for

	K				
Δ	2	5	10	15	20
-0.50	100.2	102.9	105.2	106.5	107.3
-0.30	100.8	104.8	107.8	109.4	110.3
-0.10	102.6	108.3	112.3	114.3	115.5
0.10	106.8	115.2	120.8	123.6	125.3
0.30	114.5	130.7	140.4	145.0	147.9

Maximum sample sizes

	K				
Δ	2	5	10	15	20
-0.50	100.1	102.6	104.6	105.7	106.3
-0.30	100.0	102.4	104.1	105.0	105.5
-0.10	100.0	101.9	103.2	103.9	104.3
0.10	100.0	101.1	101.8	102.2	102.4
0.30	100.0	100.3	100.5	100.6	100.7

Relative performances for F_1

	K				
Δ	2	5	10	15	20
-0.50	100.7	108.3	112.2	114.0	115.1
-0.30	100.5	107.5	110.9	112.5	113.4
-0.10	100.3	106.3	109.0	110.2	111.0
0.10	100.1	104.5	106.3	107.1	107.5
0.30	100.0	102.2	103.0	103.3	103.5

Relative performances for F_{21}

	K				
Δ	2	5	10	15	20
-0.50	102.6	127.4	136.2	140.3	142.8
-0.30	101.6	124.8	132.6	136.2	138.3
-0.10	100.7	122.0	127.7	130.6	132.3
0.10	100.3	116.4	121.2	123.2	124.4
0.30	100.1	108.9	113.2	114.4	114.9

Relative performances for F_{31}

	K				
Δ	2	5	10	15	20
-0.50	100.9	109.3	113.1	114.8	115.9
-0.30	100.5	108.2	111.5	113.0	113.9
-0.10	100.2	106.9	109.2	110.4	111.1
0.10	100.0	104.7	106.1	106.8	107.2
0.30	100.0	101.8	102.3	102.5	102.5

Relative performances for F_4

	K				
Δ	2	5	10	15	20
-0.50	100.6	106.8	110.0	111.5	112.4
-0.30	100.4	106.0	108.8	110.1	110.8
-0.10	100.1	105.0	107.0	108.0	108.6
0.10	100.0	103.3	104.5	105.1	105.4
0.30	100.0	101.1	101.5	101.7	101.7

Relative performances for F_5

Table 4.3: *Tabulated values are* $100 \times n_K/n_{fix}$ *(top left) and relative performances for objective functions* F_1, F_{21}, F_{31}, F_4 *and* F_5 *achieved by* Δ*-family tests with type I error 0.05, power 0.90 and* K *equally spaced analyses.*

F_{21}, F_4 and F_5 are moderate to poor, with the best being for F_5 and the worst for F_4, although the relative performances for these three objective functions are similar. The best relative performances are for F_1, with good relative performance for $\Delta = -0.3$ or -0.1. For all these objective functions, the relative performance improves as Δ increases, but this is at the expense of increasing the maximum sample size. Overall, the Δ-family test is a sensible choice if the primary concern is early stopping when $\mu = \delta/2$, in which case $\Delta = -0.3$ or -0.1 would be a suitable choice for a 5 or 10 analysis design.

Comparing these results with those for symmetric tests in §3.3, the overall patterns are similar. The most striking difference is that the relative performances for $K = 2$ are no longer uniformly 100.0. As was noted in the discussion of the symmetric results, there is only one degree of freedom available in selecting a symmetric test with two groups of observations, and this is used to achieve the desired error probability. Here, we have three degrees of freedom to choose the upper and lower boundary points as the two points are equal at the last analysis. Placing the boundaries to satisfy the unequal type I and type II error probabilities uses two of these degrees of freedom, thus there is no unique asymmetric two-analysis test with the desired error rates. Apart from this, the other main difference between the symmetric and asymmetric cases is that here we have a slightly different recommendation for Δ.

4.5 Performance of the error spending method

In this section, we consider the efficiency of asymmetric error spending tests, relative to the optimal tests found by backwards induction. In forming an asymmetric error spending test, we use two error spending functions which are equal in form but not in scale to determine the error spent at each analysis. If we wish to have a test with type I and type II error rates α and β respectively, and wish to use a γ-family test with

$\gamma = -2.0$, we would use error spending functions α^* and β^* where

$$\begin{aligned} \alpha^*(n_i/n_K) &= \alpha\left(1 - e^{2(n_i/n_K)}\right)\left(1 - e^2\right) \\ \text{and} \quad \beta^*(n_i/n_K) &= \beta\left(1 - e^{2(n_i/n_K)}\right)\left(1 - e^2\right). \end{aligned}$$

A test could be designed using error spending functions with different values of the parameter γ or ρ to determine the type I and type II error probabilities spent at each analysis, but this has not been considered here. Nor has the case where the two error spending functions used are from the two different families that we consider.

All tests in this section are of $H_0{:}\mu \leq 0$ against $H_1{:}\mu > 0$ with type I error probability 0.05 and power 0.90 fixed at $\mu = 0$ and $\mu = \delta$ respectively, with K equally spaced analyses. Maximum sample sizes are given as percentages of n_{fix} and relative performances are the achieved values of the relevant objective functions given as percentages of the optimal values over all tests with the same error rates and the same sequence of group sizes.

4.5.1 The γ-family of error spending tests

Table 4.4 shows maximum sample sizes and relative performances for tests using the γ-family of error spending functions, with γ taking values $-4.0, -3.0, -2.0, -1.0$, and 0.0. Maximum sample sizes are good for lower values of γ and K, remaining below 120% of n_{fix} for $\gamma \leq -2.0$ for all values of K considered.

The best relative performance values for the γ-family tests are for objective function F_1, which are within 2.5% of the optimal values in all cases considered. The relative performance for F_1 also increases as γ is lowered, so the γ-family tests with the lowest maximum sample sizes are also those with the best relative performance for F_1. Relative performance for all other objective functions improve as γ is increased, although in

	K				
γ	2	5	10	15	20
-4.0	101.4	103.8	105.2	105.9	106.2
-3.0	102.7	106.3	108.3	109.0	109.5
-2.0	105.0	110.4	113.0	113.9	114.4
-1.0	108.7	116.7	120.0	121.3	121.9
0.0	114.3	125.7	130.2	131.8	132.6

Maximum sample sizes

	K				
γ	2	5	10	15	20
-4.0	100.3	100.7	101.3	101.6	101.8
-3.0	100.3	100.8	101.3	101.6	101.8
-2.0	100.3	100.8	101.4	101.6	101.8
-1.0	100.2	100.8	101.4	101.8	101.9
0.0	100.2	100.8	101.6	102.1	102.3

Relative performances for F_1

	K				
γ	2	5	10	15	20
-4.0	100.1	101.3	102.6	103.1	103.5
-3.0	100.1	101.2	102.3	102.8	103.1
-2.0	100.1	101.0	102.0	102.4	102.7
-1.0	100.0	100.9	101.7	102.1	102.4
0.0	100.0	100.7	101.4	101.8	102.1

Relative performances for F_{21}

	K				
γ	2	5	10	15	20
-4.0	100.0	105.2	107.7	108.9	109.6
-3.0	100.0	105.0	107.0	108.1	108.6
-2.0	100.0	104.8	106.4	107.3	107.8
-1.0	100.0	104.6	106.0	106.6	107.1
0.0	100.0	104.3	105.7	106.2	106.6

Relative performances for F_{31}

	K				
γ	2	5	10	15	20
-4.0	100.1	101.2	102.4	102.9	103.2
-3.0	100.1	101.1	102.0	102.5	102.7
-2.0	100.1	100.9	101.7	102.1	102.3
-1.0	100.1	100.7	101.3	101.7	101.9
0.0	100.0	100.5	101.0	101.3	101.6

Relative performances for F_4

	K				
γ	2	5	10	15	20
-4.0	100.2	100.9	101.9	102.4	102.6
-3.0	100.1	100.8	101.7	102.1	102.3
-2.0	100.1	100.7	101.4	101.8	102.0
-1.0	100.1	100.5	101.2	101.5	101.8
0.0	100.1	100.4	101.0	101.4	101.6

Relative performances for F_5

Table 4.4: *Tabulated values are* $100 \times n_K/n_{fix}$ *(top left) and relative performances for objective functions* F_1, F_{21}, F_{31}, F_4 *and* F_5 *achieved by error spending tests using the* γ*-family of error spending function with type I error 0.05, power 0.90 and* K *equally spaced analyses.*

most cases the improvement is slow. The relative performances for F_{21}, F_4 and F_5 are all good, especially those for F_5, while the values for F_{31} are good to moderate. Overall, the tests using the γ-family of error spending functions show very good relative efficiency and maximum sample size properties, producing highly efficient tests. A test with $\gamma = -2.0$ or -3.0 would be a good test, or possibly using $\gamma = -4.0$ if the main concern was the behaviour of the test when $\mu = \delta/2$.

This behaviour is mostly unchanged from the symmetric case, with both maximum sample sizes and relative efficiency figures being very similar in both cases. The greatest difference is found by comparing the relative performance values for F_3 in table 3.4 to the corresponding values for F_{31} in table 4.4, and even in this case, the difference in relative performance values do not exceed approximately 2% of the relevant optimal values. Recommendations for suitable choices of γ are the same as in the symmetric case, $\gamma = -2.0$ or -3.0.

4.5.2 The ρ-family of error spending tests

Table 4.5 shows relative performance and maximum sample size values for tests defined by the ρ-family of error spending functions for a range of values of ρ and maximum numbers of analyses K. Calculations for other vales of ρ and K continue the patterns seen in this table.

The maximum sample size values increase as ρ and K increase, with a reasonable maximum sample size for all K considered if $\rho > 2.0$ and also if $\rho = 1.0$ when $K = 2$. This pattern is similar to that observed in the symmetric case, although here the maximum sample sizes are slightly larger.

Relative performance is worst for F_{31}, with values which are moderate to poor. Relative performances for objective functions F_{21}, F_4, and F_5 are of similar magnitude and are all good to fair. The best relative performance values are for F_1, with the ρ-family

ρ	K = 2	5	10	15	20
1.0	114.3	125.7	130.2	131.8	132.6
2.0	104.4	110.0	112.6	113.5	114.0
3.0	101.5	104.8	106.6	107.3	107.7
4.0	100.5	102.5	103.8	104.4	104.8

Maximum sample sizes

ρ	K = 2	5	10	15	20
1.0	100.2	100.8	101.6	102.0	102.3
2.0	100.3	100.6	100.9	101.1	101.2
3.0	100.3	100.9	101.4	101.6	101.7
4.0	100.3	101.0	101.7	102.1	102.3

Relative performances for F_1

ρ	K = 2	5	10	15	20
1.0	100.0	100.7	101.4	101.8	102.1
2.0	100.1	101.9	102.5	102.7	102.9
3.0	100.1	102.6	103.7	104.2	104.4
4.0	100.2	103.0	104.6	105.3	105.7

Relative performances for F_{21}

ρ	K = 2	5	10	15	20
1.0	100.0	104.3	105.7	106.2	106.6
2.0	100.0	108.4	109.7	110.2	110.6
3.0	100.0	110.5	112.7	113.6	114.2
4.0	100.0	111.5	114.9	116.3	117.0

Relative performances for F_{31}

ρ	K = 2	5	10	15	20
1.0	100.0	100.5	101.0	101.3	101.6
2.0	100.1	101.9	102.3	102.5	102.6
3.0	100.1	102.8	103.7	104.1	104.3
4.0	100.1	103.3	104.8	105.4	105.8

Relative performances for F_4

ρ	K = 2	5	10	15	20
1.0	100.1	100.4	101.0	101.4	101.6
2.0	100.1	101.3	101.7	101.9	102.0
3.0	100.1	102.0	102.8	103.1	103.3
4.0	100.2	102.3	103.6	104.1	104.4

Relative performances for F_5

Table 4.5: *Tabulated values are* $100 \times n_K/n_{fix}$ *(top left) and relative performances for objective functions* $F_1, F_{21}, F_{31}.F_4$ *and* F_5 *achieved by error spending tests using the* ρ*-family of error spending function with type I error 0.05, power 0.90 and* K *equally spaced analyses.*

tests proving highly efficient here. In all cases, the relative performance improves as ρ decreases, so a choice of a value for ρ must balance concerns of restraining the maximum sample size while achieving good efficiency. Overall, there is a combination of good maximum sample size and relative performance indicating a family of efficient tests. A choice of $\rho = 2.0$ or 3.0 would give a test with good properties. Again, this is very similar to the symmetric case.

4.6 Comparing the Δ-family, γ-family and ρ-family tests

Figure 4-3 compares the achieved values of objective functions F_1, F_{21}, F_{31}, F_4 and F_5 for Δ-family, γ-family and ρ-family tests, with optimal values of each of these objective functions. Plotted values are of achieved objective function against maximum sample size, with both values being given as percentages of the fixed sample size n_{fix}. The plotted lines join points representing tests with specific parameter values; the Δ-family tests are for $\Delta = -0.5, -0.45, \ldots, 0.25$, the ρ-family tests have $\rho = 0.8, 1.0, \ldots, 4.0$ and the γ-family tests have parameter values $\gamma = -4.0, -3.5, \ldots, 0.5$. Recall from the definitions of these tests in chapter 2 that the parameter value determines the maximum sample size as well as the shape of the boundary. All tests have size 0.05 and power 0.9, with 10 equally spaced analyses. Tests with $K = 5, 15$ and 20 show the same patterns as are seen in figure 4-3, while in the case where $K = 2$ there is little difference in performance between any of the methods.

The patterns shown are extremely similar to those observed in the symmetric case, which were shown in figure 3-3. In all cases with $n_K \leq 1.2 \times n_{fix}$, the Δ-family tests have the worst achieved objective function values, in some cases by a considerable margin. The two families of error spending function considered lead to tests which have similar performance as regards achieved objective function values. The ρ-family is slightly superior for F_1, while the γ-family is slightly superior for the other objective functions considered. Thus, the γ-family is, overall, the better choice of test design.

In practice, the Δ-family tests are much easier to design and implement, which is a considerable advantage. However, the error spending tests have the advantage of their flexibility, and figure 4-3 shows that they achieve significantly better expected sample size values. Thus, they are to be preferred in use to the Δ-family tests.

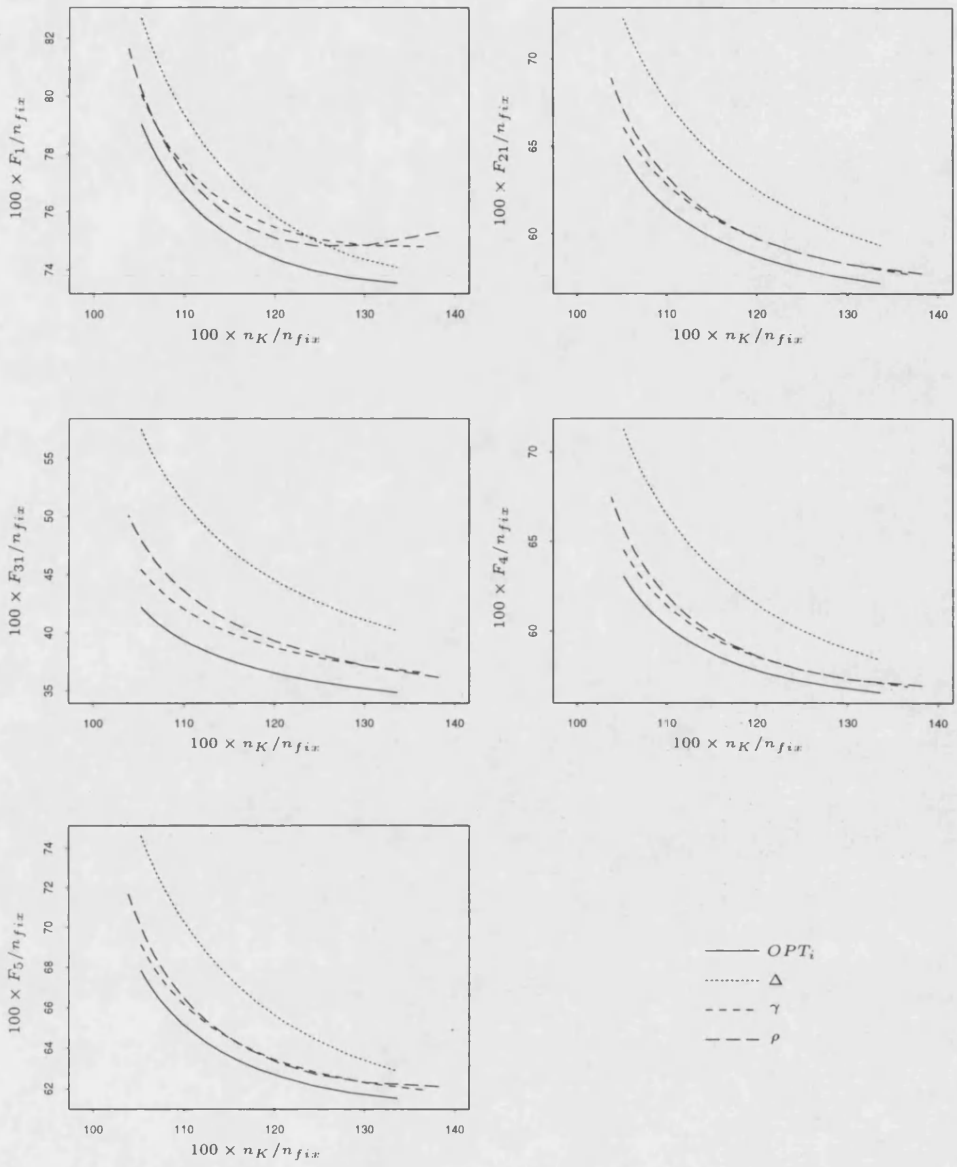


Figure 4-3: *Achieved values of F_1, F_{21}, F_{31}, F_4 and F_5 for Δ-family, γ-family and ρ-family tests. All tests have type I error 0.05 and power 0.90, with 10 equally spaced analyses. Optimal objective function values are included for comparison. Maximum sample sizes and objective function values are given as percentages of the fixed sample size n_{fix}.*

Chapter 5

Optimal group sequential tests with random group sizes

In chapters 3 and 4, we considered optimal group sequential tests when the sequence of cumulative sample sizes $n_1, n_2, \ldots, n_K$ was fixed and known in advance. We now move on to consider a situation where this is not the case and optimise group sequential tests when the sequence of sample sizes is uncertain at the design stage of the trial. We start in §5.1 by outlining a framework for optimising group sequential tests where the sequence of sample sizes is uncertain.

Once we have discussed a means of incorporating uncertainty as to the possible sample sizes that may be seen in the trial, we consider the properties of the resulting tests in §5.2. While the optimal tests are of interest in their own right, our primary motivation in developing these tests is to provide a method of assessing the performance of existing methods when the group sizes deviate from those planned. We consider the performance of the Δ-family and error spending tests in these circumstances in §5.3, using the optimal group sequential tests from §5.1 as a benchmark. Through the use of our optimal group sequential tests which incorporate group size uncertainty as a

benchmark, we can investigate the relative performance of these existing methods. This allows us to give recommendations for the design of trials where there is uncertainty as to the numbers of observations seen at each analysis, and to comment on how well the methods examined fare when the number of observations seen depart from the anticipated schedule.

Throughout this chapter, we shall refer to a sequence $n_1, n_2, \ldots, n_K$ of the total number of observations seen at each analysis as a sequence of sample sizes. We shall denote a sample size sequence by $\mathbf{n} = (n_1, \ldots, n_K)$, and we shall refer to tests designed with several possible sample size sequences in mind as random group sequential tests, as opposed to the fixed group sequential tests discussed in chapters 3 and 4.

5.1 Optimising over a sample size model

In chapters 3 and 4, we found optimal symmetric and asymmetric group sequential tests when the number of observations taken in each group was known in advance. This will not always be the case, and in this section we shall discuss a framework for introducing uncertainty as to the actual sequence of sample sizes into the group sequential trial. We do this by selecting a model for possible sample size sequences and optimising the performance of the group sequential test averaged over this sample size model.

The Bayes decision theory problem discussed in §3.1.2 and §4.2.1 needs no further modification to deal with this situation, but is restated for convenience in §5.1.2. However, the backward induction algorithm used to solve the Bayes decision theory problem must be altered to deal with the situation where we average over a sample size model, and these alterations are discussed in §5.1.3. Several possible sample size models are described in §5.1.5.

Several authors, including Pocock (1977) and Emerson & Fleming (1989), have noted the robustness of group sequential schemes to small deviations from the planned

sample size sequence. Usually, these discussions have focussed upon the achieved error probabilities, but given this robustness, it is reasonable to assume that there will be only small alterations to the expected sample size properties of a group sequential scheme if there are small perturbations from the planned sample size sequence. However, larger deviation from the planned sample size sequence can lead to greater perturbations in the achieved error probabilities and expected sample sizes, as we shall see later in this chapter.

It must be noted that it will not always be possible to describe a model for the sequence of sample sizes, much as it is not always possible to specify a fixed sequence of group sizes and adhere firmly to that schedule of analyses. However, under some circumstances it will be possible to describe the expected rate of arrival of data in such a fashion. Even if this is not the case, by considering several plausible sample size models we can investigate the performance of other group sequential schemes to help plan a clinical trial.

5.1.1 Definitions

Recall that we wish to investigate the difference in efficacy between a new treatment and a control, which may be a placebo or an existing treatment. We denote the difference in efficacy between the new and control treatments by μ, with positive values of μ indicating superiority of the new treatment. Our aim is to test the null hypothesis $H_0{:}\mu \leq 0$ against the one-sided alternative $H_1{:}\mu > 0$, with type I error probability α at $\mu = 0$ and type II error probability β at $\mu = \delta$. For the optimal group sequential tests which we shall discuss in this chapter, these error probabilities are to be averaged over a sample size model. We do not require the error probabilities of our optimal tests conditional upon a specific sample size sequence to achieve these nominal values. In later sections, we shall discuss the use of Δ-family and error spending tests with uncertain group sizes and these methods will not be required to have average error

probabilities α and β. We can take observations $X_1, X_2, \ldots$, where each observation is independently $N(\mu, \sigma^2)$ distributed, with σ^2 known and we can take a maximum of K groups of observations. For the remainder of this chapter, we use i to denote the analysis or group of observations under consideration. Unless stated otherwise, i takes values in the set $\{1, \ldots, K\}$.

Rather than fix the number of observations taken at each analysis in advance, we consider several possible sample sizes for each analysis. Define ν_i to be the number of possible values for the sample size at analysis i, then we write $n_{i,1}, \ldots, n_{i,\nu_i}$ for the possible sample sizes at analysis i. Let the sample size at analysis i be N_i, where N_i is a random variable taking values in the set $\{n_{i,1}, \ldots, n_{i,\nu_i}\}$. We define transition probabilities $\tau_k^{(0)}$ and $\tau_{j,k}^{(i)}$ as follows.

$$\tau_k^{(0)} = \mathbb{P}\{N_1 = n_{1,k}\} \qquad k = 1, \ldots, \nu_1$$

$$\tau_{j,k}^{(i)} = \mathbb{P}\{N_{i+1} = n_{i+1,k} | N_i = n_{i,j}\} \qquad \begin{cases} i = 1, \ldots, K-1 \\ j = 1, \ldots, \nu_i \\ k = 1, \ldots, \nu_{i+1} \end{cases}$$

Thus, the probability of a particular sample size sequence $n_{1,j_1}, n_{2,j_2}, \ldots, n_{K,j_K}$ occurring is

$$\mathbb{P}\{N_1 = n_{1,j_1} \cap N_2 = n_{2,j_2} \cap \ldots \cap N_K = n_{K,j_K}\} = \tau_{j_1}^{(0)} \tau_{j_1,j_2}^{(1)} \cdots \tau_{j_{K-1},j_K}^{(K-1)}. \qquad (5.1)$$

As noted above, we shall be finding group sequential tests with error probabilities α and β at $\mu = 0$ and $\mu = \delta$ respectively. However, these error probabilities are averaged over the sample size model, rather than holding for each possible sample size sequence. Any specific sample size sequence $\mathbf{n}$ will have associated error probabilities and the contribution these probabilities make to the overall error rate is weighted by the probability of the sample size sequence $\mathbf{n}$ occurring, as given by equation (5.1).

As we shall see in §5.2, the error probabilities conditional upon different sample size sequences can vary considerably.

For the rest of this chapter, we reserve the subscript j for use as an index of the possible sample sizes at analysis i. Thus, unless stated otherwise, j takes values $1, 2, \ldots, \nu_i$. As in the fixed groups case discussed in chapters 3 and 4, we define N to be the number of observations taken by the termination of the trial and we define the summary statistic of the first n observations, S_n, to be $S_n = \sum_{k=1}^{n} X_k$. Thus, $S_n \sim N(n\mu, n\sigma^2)$.

The action taken at analysis i depends upon the number of observations seen, the sum of these observations, and the number of analyses already carried out. However, our action does not depend upon the sequence of cumulative sample sizes $n_1, \ldots, n_i$ other than through the current sample size n_i. Thus, whether we stop the trial or continue to the next analysis having seen $n_{i,j}$ observations with sum $S_{n_{i,j}}$ is determined by constants $a_{i,j}$ and $b_{i,j}$, with $a_{i,j} \leq b_{i,j}$, as follows.

If	$S_{n_{i,j}} \leq a_{i,j}$	STOP, accept H_0,
if	$a_{i,j} < S_{n_{i,j}} < b_{i,j}$	continue to analysis $i+1$,
and if	$b_{i,j} \leq S_{n_{i,j}}$	STOP, reject H_0.

To ensure termination of the trial at analysis K, we set $a_{K,j} = b_{K,j}$. Thus, if we reach analysis K we shall either accept or reject H_0.

This type of random group sequential boundary is different from the fixed group sequential boundaries discussed in earlier chapters. When the sample size sequence is fixed and known in advance, there is a single continuation region (a_i, b_i) at each analysis i. Here, we have a set of different continuation regions $\{(a_{i,j}, b_{i,j})\}$ at analysis i, and which region is used is dependent upon the observed sample size at analysis i.

5.1.2 The Bayes problem

The Bayes decision theory problem used to find optimal asymmetric group sequential tests with fixed group sizes described in §4.2.1 is used without further modifications to find the optimal random group sequential test averaged over a sample size model. However, the implementation via a backward induction algorithm requires modifications, which are detailed in the following section. We restate the Bayesian decision theory problem here for convenience.

We wish to choose between possible decisions D_0: '$\mu = 0$' and D_δ: '$\mu = \delta$'. We place a prior $\pi(\mu)$ on μ and define the cost of taking one observation to be $c(\mu)$. We define $\mathcal{M}$ to be the set of values of μ upon which the prior $\pi(\mu)$ places a non-zero probability mass. We also define the loss function $L_2(D, \mu)$ which gives the cost of wrong decisions to be $L_2(D_0, \delta) = d_\delta$, $L_2(D_\delta, 0) = d_0$ and $L_2(D, \mu) = 0$ otherwise.

Our goal is to find a test minimising one of the objective functions defined on page 50. The expected cost of any decision rule is

$$\begin{aligned} \mathbb{E}\{\text{cost}\} &= \mathbb{E}\{\text{cost of sampling}\} + \mathbb{E}\{\text{cost of decision}\} \\ &= k_1 F_r + k_2 \left(d_\delta \pi(\delta) \mathbb{P}_\delta\{D_0\} + d_0 \pi(0) \mathbb{P}_0\{d_\delta\} \right), \end{aligned}$$

where F_r is the objective function we wish to minimise and the constants k_1 and k_2 are determined by the choice of objective function. If we wish to minimise F_1, $k_1 = k_2 = 1/3$, if we are minimising F_{21}, $k_1 = 1$ and $k_2 = 1/2$, if we are minimising F_{31} then $k_1 = 1/2$ and $k_2 = 1/4$, and if we wish to minimise F_4, then $k_1 = 1$ and $k_2 = 1/10$. Appropriate priors and cost of sampling functions were defined in §4.2.1 and remain unchanged.

For any given values of d_0 and d_δ, the Bayes rule will have probabilities of making incorrect decisions $\mathbb{P}_\delta\{D_0|d_0, d_\delta\}$ and $\mathbb{P}_0\{D_\delta|d_0, d_\delta\}$, and the Bayes rule will minimise

the total expected cost of all decision rules with these probabilities of wrong decisions. Hence, the Bayes rule must minimise F_r amongst all decision rules with these probabilities of wrong decisions. We then search over d_0 and d_δ to find the Bayes rule which has probabilities of wrong decisions $\mathbb{P}_\delta\{D_0|d_0, d_\delta\} = \alpha$ and $\mathbb{P}_0\{D_\delta|d_0, d_\delta\} = \beta$ and thus identify the group sequential test with the desired error probabilities α and β averaged over our sample size model.

One complication that arises from using this approach for random sample sizes occurs when calculating $\mathbb{P}_\delta\{D_0|d_0, d_\delta\}$ and $\mathbb{P}_0\{D_\delta|d_0, d_\delta\}$ for the Bayes rule specified by particular values of d_0 and d_δ. The method used in the fixed group case proves to be impractical here and so an iterative method, described in §5.1.4 is used.

5.1.3 Adapting the backward induction algorithm to random sample sizes

The overall structure of the backwards induction algorithm described in §3.1.3 and §4.2.2 remains unchanged. However, when we calculate the expected cost of continuing to the next analysis and proceeding optimally there, we must now consider the different possible sizes of the next group of observations.

The posterior probability of any value of $\mu \in \mathcal{M}$ at analysis i, given $n_{i,j}$ observations and an observed value $s_{n_{i,j}}$ of the summary statistic $S_{n_{i,j}}$, is

$$p^{(i)}(\mu|s_{n_{i,j}}) \propto \pi(\mu) f_\mu^{(i)}(s_{n_{i,j}}),$$

where $f_\mu^{(i)}$ is the probability density function of $S_{n_{i,j}}$ with each observation having mean μ;

$$f_\mu^{(i)}(s_{n_{i,j}}) = \frac{1}{\sqrt{2\pi n_{i,j}\sigma^2}} \exp\left\{\frac{(s_{n_{i,j}} - n_{i,j}\mu)^2}{-2n_{i,j}\sigma^2}\right\}.$$

We note that the posterior probability for μ is also a function of $n_{i,j}$, but suppress the

sample size in our notation for simplicity. The dependence on the sample size can be seen through the subscript of $s_{n_{i,j}}$.

We start by considering what action to take at analysis K, should the trial proceed that far. In order to ensure termination of the trial, we set the boundary points $a_{K,j} = b_{K,j}$. These points are set to the value of $S_{n_{K,j}}$ such that the expected losses of making either decision D_0 or D_δ are equal; we label this value $s^*_{n_{K,j}}$ and solve the equation

$$\begin{aligned} \mathbb{E}\left\{\text{cost of } D_0 | S_{n_{K,j}} = s^*_{n_{K,j}}\right\} &= \mathbb{E}\left\{\text{cost of } D_\delta | S_{n_{K,j}} = s^*_{n_{K,j}}\right\} \\ \Rightarrow \qquad d_\delta p^{(K)}(\delta | s^*_{n_{K,j}}) &= d_0 p^{(K)}(0 | s^*_{n_{K,j}}) \end{aligned}$$

to find that

$$s^*_{n_{K,j}} = \frac{\delta n_{K,j}}{2} - \frac{\sigma^2}{\delta} \log\left\{\frac{d_\delta \pi(\delta)}{d_0 \pi(0)}\right\} \tag{5.2}$$

For analyses $i = 1, \ldots, K-1$, we define $\gamma^{(i)}(s_{n_{i,j}})$ to be the expected cost of stopping and making a decision given sample size $n_{i,j}$ and an observed value $s_{n_{i,j}}$ of the summary statistic $S_{n_{i,j}}$. Similarly, we define $\beta^{(i)}(s_{n_{i,j}})$ to be the expected cost of continuing to the next analysis and acting optimally there. As with the posterior for μ, we suppress the sample size $n_{i,j}$ in this notation for simplicity.

At analysis i, we consider the boundary points for each possible group size $n_{i,j}$. To determine the boundary points $a_{i,j}$ and $b_{i,j}$, we first calculate $s^*_{n_{i,j}}$, using equation (5.2) with $n_{i,j}$ in place of $n_{K,j}$. If the expected cost of stopping and making a decision when $S_{n_{i,j}} = s^*_{n_{i,j}}$ is less than that of continuing to analysis $i+1$ and proceeding optimally, we shall set $a_{i,j} = b_{i,j} = s^*_{i,j}$. Otherwise, we search for two values of $S_{n_{i,j}}$ satisfying the equation $\gamma^{(i)}(s_{n_{i,j}}) = \beta^{(i)}(s_{n_{i,j}})$, with one of these values above $s^*_{n_{i,j}}$ and the other below $s^*_{n_{i,j}}$. We then set the boundary points to these values, with $a_{i,j} < s^*_{n_{i,j}} < b_{i,j}$. As in the fixed group sequential tests in chapters 3 and 4, these searches rely upon the monotonicity of $\gamma^{(i)}(s_{n_{i,j}}) - \beta^{(i)}(s_{n_{i,j}})$, which was discussed on page 29.

We now need to discuss the calculation of $\gamma^{(i)}(s_{n_{i,j}})$ and $\beta^{(i)}(s_{n_{i,j}})$. Firstly, we define $F^{(i+1)}_{n_{i+1,k}}(s_{n_{i+1,k}}|s_{n_{i,j}})$ to be the cumulative distribution function of $S_{n_{i+1,k}}$ given $n_{i+1,k}$, $n_{i,j}$ and $s_{n_{i,j}}$, again suppressing the term $n_{i,j}$ in this notation for simplicity. Thus,

$$\begin{aligned} & dF^{(i+1)}_{n_{i+1,k}}(s_{n_{i+1,k}}|s_{n_{i,j}}) \\ = \; & g^{(i+1)}(s_{n_{i+1,k}}|s_{n_{i,j}})ds_{n_{i+1,k}} \\ = \; & \sum_{\mu\in\mathcal{M}} \left\{p^{(i)}(\mu|s_{n_{i,j}})h^{(i+1)}_{\mu}(s_{n_{i+1,k}}|n_{i+1,k}, s_{n_{i,j}})\right\} ds_{n_{i+1,k}}, \end{aligned}$$

where g is the density function of $S_{n_{i+1,k}}$ given $n_{i+1,k}, n_{i,j}$ and $s_{n_{i,j}}$ and h is the density function of $S_{n_{i+1,k}}$ given $n_{i+1,k}, n_{i,j}, s_{n_{i,j}}$ and μ.

The expected cost of stopping and making a decision at analysis i with sample size $n_{i,j}$ and observed value $s_{n_{i,j}}$ of the summary statistic $S_{n_{i,j}}$, is

$$\gamma^{(i)}(s_{n_{i,j}}) = \min\left\{d_{\delta}p^{(i)}(\delta|s_{n_{i,j}}), d_0 p^{(i)}(0|s_{n_{i,j}})\right\},$$

which is easy to calculate once the posterior probabilities of $\mu = 0$ and $\mu = \delta$ are known. Calculating the expected cost of continuing to the next analysis and proceeding optimally there is more complicated, as it is necessary to consider the different possible values of N_{i+1}. The expected cost of continuing to the next analysis and acting optimally there is given by

$$\begin{aligned} \beta^{(i)}(s_{n_{i,j}}) \; = \; & \mathbb{E}\{\text{cost of observing group } i+1\} \; + \\ & \mathbb{E}\{\text{cost of optimal action at analysis } i+1\}. \end{aligned}$$

For analysis i, $i = 1, \ldots, K-1$, the expected cost of observing group $i+1$ when the current sample size is $n_{i,j}$ is

$$\sum_{\mu\in\mathcal{M}} \left[c(\mu)p^{(i)}(\mu|s_{n_{i,j}}) \left\{\sum_{k=1}^{\nu_{i+1}} \tau^{(i)}_{j,k}(n_{i+1,k} - n_{i,j})\right\}\right].$$

If we continue from analysis $K-1$, the expected cost of acting optimally at analysis K, given the current sample size $n_{K-1,j}$ and an observed value $s_{n_{K-1,j}}$ of $S_{n_{K-1,j}}$ is

$$\sum_{k=1}^{\nu_K} \left\{ \tau_{j,k}^{(K-1)} \int_{\mathbb{R}} \gamma^{(K)}(s_{n_{K,k}}) dF_{n_{K,k}}^{(K)}(s_{n_{K,k}}|s_{n_{K-1,j}}) \right\}.$$

For analyses $i = 1, \ldots, K-2$, we must also consider the possibility of continuing past analysis $i+1$. Thus, if we continue to analysis $i+1$, the expected cost of acting optimally there, given the current sample size $n_{i,j}$ and observed value $s_{n_{i,j}}$ of $S_{n_{i,j}}$ is given by

$$\sum_{k=1}^{\nu_{i+1}} \left\{ \tau_{j,k}^{(i)} \int_{\mathbb{R}} \min\left\{ \gamma^{(i+1)}(s_{n_{i+1,k}}), \beta^{(i+1)}(s_{n_{i+1,k}}) \right\} dF_{n_{i+1,k}}^{(i+1)}(s_{n_{i+1,k}}|s_{n_{i,j}}) \right\}.$$

Hence, the expected cost of proceeding optimally from analysis i ($i = 1, \ldots, K-2$) after $n_{i,j}$ observations with sum $s_{n_{i,j}}$ is

$$\begin{aligned} \beta^{(i)}(s_{n_{i,j}}) \;=\; & \sum_{\mu \in \mathcal{M}} \left[c(\mu) p^{(i)}(\mu|s_{n_{i,j}}) \left\{ \sum_{k=1}^{\nu_{i+1}} \tau_{j,k}^{(i)}(n_{i+1,k} - n_{i,j}) \right\} \right] \;+ \\ & \sum_{k=1}^{\nu_{i+1}} \left\{ \tau_{j,k}^{(i)} \int_{\mathbb{R}} \min\left\{ \gamma^{(i+1)}(s_{n_{i+1,k}}), \beta^{(i+1)}(s_{n_{i+1,k}}) \right\} dF_{n_{i+1,k}}^{(i+1)}(s_{n_{i+1,k}}|s_{n_{i,j}}) \right\}, \end{aligned} \tag{5.3}$$

and the expected cost of proceeding optimally from analysis $K-1$ after $n_{K-1,j}$ observations with sum $s_{n_{K-1,j}}$ is

$$\begin{aligned} \beta^{(K-1)}(s_{n_{K-1,j}}) \;=\; & \sum_{\mu \in \mathcal{M}} \left[c(\mu) p^{(K-1)}(\mu|s_{n_{K-1,j}}) \left\{ \sum_{k=1}^{\nu_K} \tau_{j,k}^{(K-1)}(n_{K,k} - n_{K-1,j}) \right\} \right] \\ & + \sum_{k=1}^{\nu_K} \left\{ \tau_{j,k}^{(K-1)} \int_{\mathbb{R}} \gamma^{(K)}(S_{n_{K,k}}) dF_{n_{K,k}}^{(K)}(S_{n_{K,k}}|s_{n_{K-1,j}}) \right\}. \end{aligned}$$

These equations are the same as equations (3.1) and (3.2) in the fixed groups case, with the addition of the sums over the possible sample sizes at analysis $i+1$.

The integral of $\beta^{(i+1)}(s_{n_{i+1,k}})$ in equation (5.3) is calculated numerically, requiring

the evaluation of $\beta^{(i+1)}(n_{i+1,k}, s_{n_{i+1,k}})$ on a grid of points in the region $[a_{i+1,k}, b_{i+1,k}]$. These values are evaluated for each of the sample sizes $n_{i+1,k}, \ldots, n_{i+1,\nu_{i+1}}$ immediately after finding the boundary points $\{a_{i+1,k}, b_{i+1,k}; k = 1, \ldots, \nu_{i+1}\}$, before considering the action to be taken at analysis i. Thus, we start by considering the optimal course of action at analysis K, then proceed backwards to analysis $K-1$, $K-2$, and so on to analysis 1. This is summarised in algorithm 3.

Algorithm 3: Asymmetric random groups algorithm

ANALYSIS K

- For $j = 1, \ldots, \nu_K$:
  - Find $s^*_{n_{K,j}}$ such that $d_\delta p^{(K)}(\delta|s^*_{n_{K,j}}) = d_0 p^{(K)}(0|s^*_{n_{K,j}})$.
  - Set $a_{K,j} = b_{K,j} = s^*_{n_{K,j}}$.

ANALYSES $K-1, \ldots, 2$

- For $i = K-1, \ldots, 2$:
  - For $j = 1, \ldots, \nu_i$:
    - Find $s^*_{n_{i,j}}$ such that $d_\delta p^{(i)}(\delta|s^*_{n_{i,j}}) = d_0 p^{(i)}(0|s^*_{n_{i,}})$.
    - If $\gamma^{(i)}(s^*_{n_{i,j}}) > \beta^{(i)}(s^*_{n_{i,j}})$:
      - Find $a_{i,j}$ such that $\gamma^{(i)}(a_{i,j}) = \beta^{(i)}(a_{i,j})$, with $a_{i,j} < s^*_{n_{i,j}}$.
      - Find $b_{i,j}$ such that $\gamma^{(i)}(b_{i,j}) = \beta^{(i)}(b_{i,j})$, with $b_{i,j} > s^*_{n_{i,j}}$.
      - Evaluate $\beta^{(i)}(s_{n_{i,j}})$ on a grid of values of $s_{n_{i,j}}$ from $a_{i,j}$ to $b_{i,j}$.
    - Otherwise, set $a_{i,j} = b_{i,j} = s^*_{n_{i,j}}$.

ANALYSIS 1

- For $j = 1, \ldots, \nu_1$:
  - Find $s^*_{n_{1,j}}$ such that $d_\delta p^{(1)}(\delta|s^*_{n_{1,j}}) = d_0 p^{(1)}(0|s^*_{n_{1,j}})$.

- If $\gamma^{(1)}(s^*_{n_{1,j}}) > \beta^{(1)}(s^*_{n_{1,j}})$:
  - Find $a_{1,j}$ such that $\gamma^{(1)}(a_{1,j}) = \beta^{(1)}(a_{1,j})$, with $a_{1,j} < s^*_{n_{1,j}}$.
  - Find $b_{1,j}$ such that $\gamma^{(1)}(b_{1,j}) = \beta^{(1)}(b_{1,j})$, with $b_{1,j} > s^*_{n_{1,j}}$.
- Otherwise, set $a_{1,j} = b_{1,j} = s^*_{n_{1,j}}$.

Comparing this algorithm to those for the symmetric and asymmetric fixed-groups problems on pages 29 and 55, it can clearly be seen that this algorithm is fundamentally similar to those algorithms, with the added complication of considering the different possible sample sizes at the next analysis.

5.1.4 Iterative error probability and objective function evaluation

Extending the method of finding optimal group sequential tests by means of a Bayes decision theory problem to finding group sequential tests averaged over a model for possible sample size sequences creates complications when calculating the error probabilities of a specific boundary. In the search for a Bayes rule with the desired probabilities of making a wrong decision, the probabilities of making each of the two possible incorrect decisions must be calculated for each pair of values of d_0 and d_δ considered. In the fixed groups case, this was achieved by finding the group sequential boundary implied by the Bayesian decision theory problem with particular values of d_0 and d_δ via the backward induction algorithm and then using numerical integration of the joint distribution of the summary statistics $\{S_{n_1}, \ldots, S_{n_K}\}$ to determine the probabilities of falsely accepting and rejecting H_0 under $\mu = \delta$ and $\mu = 0$ respectively. An efficient algorithm for these calculations is given by Jennison (1994). However, this approach only calculates the error probabilities for a specific sequence of sample sizes $\mathbf{n}$. With the introduction of a model for this sequence, calculating the error probabilities for each possible sequence of sample sizes and weighting the resulting error probabilities by the product of the transition probabilities for the sample size sequence in question,

given in equation (5.1), would be inefficient. Instead, we find the error probabilities by means of an iterative calculation. Define the probabilities of incorrectly rejecting or accepting H_0 if we continue from analysis i, when we have seen $n_{i,j}$ observations yielding a summary statistic value of $s_{n_{i,j}}$ at analysis i to be respectively

$$\begin{aligned} & \psi_u^{(i)}(s_{n_{i,j}}) = \mathbb{P}_0\{\text{reject } H_0 \text{ at analysis } i+1 \text{ or later}|s_{n_{i,j}}\} \\ \text{and} \quad & \psi_l^{(i)}(s_{n_{i,j}}) = \mathbb{P}_\delta\{\text{accept } H_0 \text{ at analysis } i+1 \text{ or later}|s_{n_{i,j}}\}. \end{aligned}$$

Then at analysis $K-1$,

$$\psi_u^{(K-1)}(s_{n_{K-1,j}}) = \sum_{k=1}^{\nu_{K-1}} \tau_{j,k}^{(K-1)} \mathbb{P}_0\{s_{n_{K,k}} > b_{K,k}|s_{n_{K-1,j}}\},$$

and at analysis i, for $i = 1, \ldots, K-2$,

$$\psi_u^{(i)}(s_{n_{i,j}}) = \sum_{k=1}^{\nu_{i+1}} \tau_{j,k}^{(i)} \left[\begin{array}{c} \mathbb{P}_0\{S_{n_{i+1,k}} > b_{i+1,k}|s_{n_{i,j}}\} \quad + \\ \int_{a_{i+1,k}}^{b_{i+1,k}} \psi_u^{(i+1)}(s_{n_{i+1,k}}) dF_{n_{i+1,k}}^{(i+1)}(s_{n_{i+1,k}}|s_{n_{i,j}}) \end{array} \right] \tag{5.4}$$

with similar iterative equations for $\psi_l^{(i)}(s_{n_{i,j}})$. At each analysis i, and for each possible sample size $n_{i,j}$, these functions can be evaluated on a grid of values of $s_{n_{i,j}}$ in order to numerically evaluate the integral in equation (5.4) when calculating $\psi_u^{(i-1)}$ and $\psi_l^{(i-1)}$. The grid of points used is the same as the grid upon which $\beta^{(i)}(s_{n_{i,j}})$ is evaluated and these calculations are carried out concurrently with the backwards induction algorithm used to find the optimal random group sequential test. The probabilities of falsely rejecting and accepting H_0 are then $\psi_u^{(0)}(0,0)$ and $\psi_l^{(0)}(0,0)$ respectively.

A similar method can be used to calculate the expected sample sizes under different values of μ, by defining $\omega_\mu^{(i)}(s_{n_{i,j}})$ to be the expected additional number of observations to be taken after continuing from analysis i, given a sample size $n_{i,j}$ at analysis i, with an observed summary statistic value $s_{n_{i,j}}$.

5.1.5 Some sample size models

Several models for the possible sample sizes will be considered in the following discussions. In this section, we introduce and describe these models. They are defined by the maximum number of analyses, the possible sample sizes at each analysis and the transition probabilities between possible sample sizes at the current and subsequent analyses. In contrast to the discussions in chapters 3 and 4, we also specify the values of δ and σ^2 as the sample sizes are given as actual numbers of observations rather than percentages of the fixed sample size n_{fix}. It would be perfectly feasible to specify these sample sizes relative to n_{fix}, thus making these results invariant to δ and σ^2, but the tests are more intuitively understandable when discussed in terms of the numbers of observations. We note that, as discussed on page 4, non-integer sample sizes have a valid interpretation in generalising these results to trials which have more complicated statistical formulations than our simple sum of independent and identically distributed normal random variables. However, all the examples discussed in this chapter have integer sample sizes.

For all the models we consider, we are testing H_0:$\mu \leq 0$ against H_1:$\mu > 0$ with type I and type II error probabilities α and β set at $\mu = 0$ and $\mu = \delta$ respectively. Recall that these error probabilities are averaged over the sample size model, and that the error probabilities conditional upon the observed sample size sequence are allowed to vary. Observations are independently $N(\mu, \sigma^2)$ distributed, with σ^2 known. For the examples studied in this chapter, $\delta = 0.25, \sigma = 1.0$ and the error probabilities are $\alpha = \beta = 0.05$. These values lead to a required sample size of 43.30 (to 2 decimal places) if no interim analyses are to be carried out.

For models 1, 2, and 3, the maximum number of analyses is is three. We also consider three models (models 4, 5, and 6) which permit a maximum of five analyses. Of these, model 4 is similar in spirit to model 1, model 4 is similar to model 2, and model 6 is

similar in spirit to model 1.

Model 1

For this model, the possible sample sizes at analysis i $(i = 1, 2, 3)$ are

$$N_i \in \{15i, 15i \pm 4, 15i \pm 6\}.$$

The probabilities of each sample size occurring at any given analysis are in the ratio 1:2:3:2:1. More explicitly, the possible sample sizes are

$$\begin{aligned} N_1 &\in \{9, 11, 15, 19, 21\}, \\ N_2 &\in \{24, 26, 30, 34, 36\}, \\ \text{and} \qquad N_3 &\in \{39, 41, 45, 49, 51\}, \end{aligned}$$

and transition probabilities are

$$\left.\begin{array}{lllll} \tau_1^{(0)} = \tau_5^{(0)} = 1/9 & & \tau_{j,1}^{(i)} = \tau_{j,5}^{(i)} = 1/9 & \\ \tau_2^{(0)} = \tau_4^{(0)} = 2/9 & & \tau_{j,2}^{(i)} = \tau_{j,4}^{(i)} = 2/9 & \\ \tau_3^{(0)} = 1/3 & & \tau_{j,3}^{(i)} = 1/3 & \end{array}\right\} \begin{array}{l} i = 1, 2 \\ j = 1, \ldots, 5. \end{array}$$

For this relatively simple model, it is easier to list the sample sizes and transition probabilities explicitly. However, the other models we shall consider are more complex and it will be simpler to describe those models in the more compact form used above.

This model represents a situation where there is the same degree of uncertainty at each analysis and there is no sample size which could occur at more than one analysis. The total sample size at earlier analyses has no effect upon the total sample size at later analyses, although the early total sample sizes do affect the group sizes which occur

later in the trial. This would be appropriate if we could reschedule later analyses after seeing the rate at which observations are accrued early in the trial. While rescheduling later analyses in light of observed data will violate the design of any group sequential test, we are allowed to alter the schedule of analyses to allow for quicker or slower accrual of data than was anticipated.

Model 2

For this model, the possible sample sizes at analysis i are

$$N_i \in \{15i, 15i \pm 2, 15i \pm 4, 15i \pm 6, 15i \pm 8, 15i \pm 10\}.$$

In most cases, the probabilities of the possible sample sizes at the next analysis are in the ratio 1:2:3:4:5:6:5:4:3:2:1, but there are situations where some of the possible sample sizes at the next analysis are not feasible since they have already been reached or even exceeded. In this case, the probabilities of these impossible group sizes occurring are set to zero and the remaining transition probabilities re-normalised to sum to one. For example, if we see $n_1 = 25$, the greatest possible size for the first group, N_2 cannot take values 20, 22, or 24. We then set $\tau^{(1)}_{11,1} = \tau^{(1)}_{11,2} = \tau^{(1)}_{11,3} = 0$, and the remaining transition probabilities for going from $n_1 = 25$ to $N_2 = n_2$ are set in the ratio 4:5:6:5:4:3:2:1.

As with model 1, there is the same degree of uncertainty as to the possible sample size at each analysis. However, in this model the degree of uncertainty is much greater. Apart from prohibiting impossible group sizes and re-normalising the remaining transition probabilities, the total sample size at earlier analyses does not affect the size of the next group of observations.

Model 3

This is the third three-analysis model considered and is different in style to models 1 and 2. In this model, whatever the current sample size, we have the possibility of there being 12, 15 or 18 observations in the next group. In each case, the probability of the next group consisting of 12 observations is 0.25, the probability of seeing 15 observations in the next group is 0.5 and the probability of seeing 18 observations is 0.25. This results in possible sample sizes

$$\begin{aligned} & N_1 \in \{12, 15, 18\}, \\ & N_2 \in \{24, 27, 30, 33, 36\}, \\ \text{and} \quad & N_3 \in \{36, 39, 42, 45, 48, 51, 54\} \end{aligned}$$

and transition probabilities

$$\begin{array}{lccc} & \tau_1^{(0)} = 0.25 & \tau_2^{(0)} = 0.5 & \tau_3^{(0)} = 0.25 \\ \text{and} & \tau_{j,j}^{(i)} = 0.25 & \tau_{j,j+1}^{(i)} = 0.5 & \tau_{j,j+2}^{(i)} = 0.25 \end{array}$$

for $i = 1, 2$ and $j = 1, \ldots, \nu_i$. Unlike models 1 and 2, the size of early groups of observations strongly affects the possible sample sizes at later analyses, but not the subsequent group sizes.

Model 4

This is the first five-analysis model we consider. For this model, possible sample size values at analysis i are

$$N_i \in \{10i, 10i \pm 3, 10i \pm 5\}.$$

Transition probabilities are in the ratio 1:2:3:2:1, except for the largest possible sample size at each analysis, when $\tau_{5,1}^{(i)} = 0$, and the remaining transition probabilities are in

the ratio 2:3:2:1. This model is similar in sprit to model 1, but with a greater number of analyses allowed.

Model 5

This model has the greatest number of possible sample sizes of all the models considered in this chapter. There are a maximum of five analyses allowed, and for each analysis i the possible sample sizes are

$$N_i \in \{10i, 10i \pm 1, 10i \pm 2, 10i \pm 4, 10i \pm 6, 10i \pm 8\}.$$

There are eleven possible sample sizes at each analysis, and the probabilities of each sample size occurring are usually in the ratio 1:2:3:4:5:6:5:4:3:2:1. As for models 2 and 4, in the case of impossible transitions (for example going from $n_1 = 18$ to $n_2 = 16$), the relevant transition probabilities are set to zero and the remaining values re-normalised.

Model 6

In this model, whatever the current sample size, we have the possibility of there being 6, 9 or 12 observations in the next group, with probability 0.25, 0.5 and 0.25 respectively. This results in the possible sample sizes at each analysis being

$$\begin{aligned} & N_1 \in \{6, 9, 12\}, \\ & N_2 \in \{12, 15, 18, 21, 24\}, \\ & N_3 \in \{18, 21, 24, 27, 30, 33, 36\}, \\ & N_4 \in \{24, 27, 30, 33, 36, 39, 42, 45, 48\}, \\ \text{and} \quad & N_5 \in \{30, 33, 36, 39, 42, 45, 48, 51, 54, 57, 60\}. \end{aligned}$$

The transition probabilities for this model are

$$\tau_1^{(0)} = 0.25 \qquad \tau_2^{(0)} = 0.5 \qquad \tau_3^{(0)} = 0.25$$

and $$\tau_{j,j}^{(i)} = 0.25 \qquad \tau_{j,j+1}^{(i)} = 0.5 \qquad \tau_{j,j+2}^{(i)} = 0.25$$

for $i = 1, \ldots, 4$ and $j = 1, \ldots, \nu_i$.

In models 3 and 6, the sample size at later analyses depends greatly upon the size of earlier groups, while in models 1, 2, 4, and 5 this dependence is much weaker, being limited to the prohibiting of impossible transitions. Models 1, 2, 4, and 5 would be suitable for a trial where the calendar time of future analyses could be rescheduled in light of the rate of accrual, while models 3 and 6 represent a situation where the calendar time between analyses is fixed and thus lower than anticipated accrual at early analyses would reduce the final sample size possible. This results in greater uncertainty as to the final sample size than about the sample size at earlier analyses.

5.1.6 An example of a random group sequential test

Figure 5-1 shows the boundaries of an optimal random group sequential test. The boundaries are designed to test H_0:$\mu \leq 0$ against H_1:$\mu > 0$ with type I error probability $\alpha = 0.05$ at $\mu = 0$ and type II error probability $\beta = 0.05$ at $\mu = 0.25$. The test follows sample size model 5 and is optimised for $F_1 = \mathbb{E}_{\mu=0.125}\{N\}$. Sample size model 5 is based upon an anticipated sample size path $\mathbf{n} = (10, 20, 30, 40, 50)$, with a large degree of possible variation about this anticipated schedule of analyses. The boundaries show how several different continuation regions are possible at each analysis, and also how the continuation region after a particular number of observations may vary depending upon how many analyses have been taken before reaching that point. In this example, a total of 16 observations could be seen at the first analysis or after two analyses have been carried out.

Figure 5-1 also shows a hypothetical path of observed data, which has resulted in

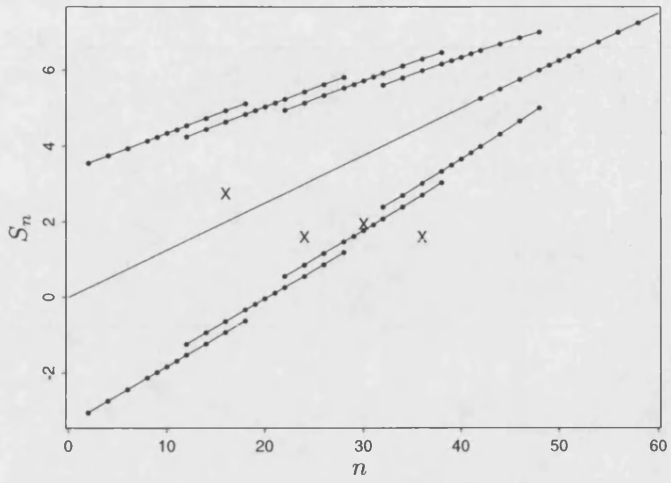


Figure 5-1: *An optimal random group sequential test. The test follows sample size model 5 and is optimised for objective function F_1. Overall type I and type II error probabilities are 0.05.*

the trial being terminated after 4 analyses with the conclusion to accept H_0. This hypothetical trial has seen a large initial group of 16 observations, followed by three smaller groups of 8, 6, and 6 observations.

5.2 Conditional performance of the optimal random group sequential tests

Using the method discussed in §5.1, we can find optimal group sequential tests averaged over sample size models which minimise the objective functions we defined on page 50. We refer to these tests as being random group sequential tests, as opposed to fixed group sequential tests as discussed in chapters 3 and 4. Several candidate sample size models were described in §5.1.5. The resulting tests have a set of boundary values for use at each analysis; which pair of values are used for the upper and lower bounds of the continuation region at analysis i will depend upon the observed value of the sample

size N_i. In practice, a realisation of the sample size sequence $\mathbf{n} = (n_{1,j_1}, \ldots, n_{K,j_K})$ would be observed, leading to a fixed group sequential test with a single continuation region (a_{i,j_i}, b_{i,j_i}) at each analysis i. Thus, the random group sequential test consists of a set of fixed group sequential tests.

A major motivation for developing this method of calculating optimal random group sequential tests is to provide a benchmark to assess the performance of existing methods, such as the Δ-family and error spending tests we have discussed in earlier chapters. However, we also wish to examine the properties of our optimal random group sequential tests in their own right. They are unusual, in frequentist terms, in that they permit variation in type I error probability, depending upon the observed sample size sequence. We shall consider the properties of our random group sequential tests conditional upon the sample size sequence observed. Firstly, we consider the achieved error probabilities in §5.2.1, then move on to consider the expected sample sizes of these tests in §5.2.2. We also consider the conditional performance of the corresponding decision rule from a Bayesian perspective in §5.2.3.

5.2.1 Achieved error probabilities

We define $\widetilde{\alpha}(\mathbf{n})$ to be the achieved type I error probability of the random group sequential test with sample size sequence $\mathbf{n} = (n_1, \ldots, n_K)$, and similarly we define $\widetilde{\beta}(\mathbf{n})$ to be the achieved type II error probability for the sample sample size sequence. Any random group sequential test will have a range of values of $\widetilde{\alpha}(\mathbf{n})$ and $\widetilde{\beta}(\mathbf{n})$, corresponding to the different possible sample size sequences.

Figure 5-2 shows values of $\widetilde{\alpha}(\mathbf{n})$ achieved by a symmetric test with $\alpha = \beta = 0.05$ using sample size model 1. Recall that this model permits a maximum of $K = 3$ analyses, and there is no overlap between the sets of possible sample sizes at consecutive analyses. The test is optimised for F_1, the expected sample size when $\mu = \delta/2$. Since the

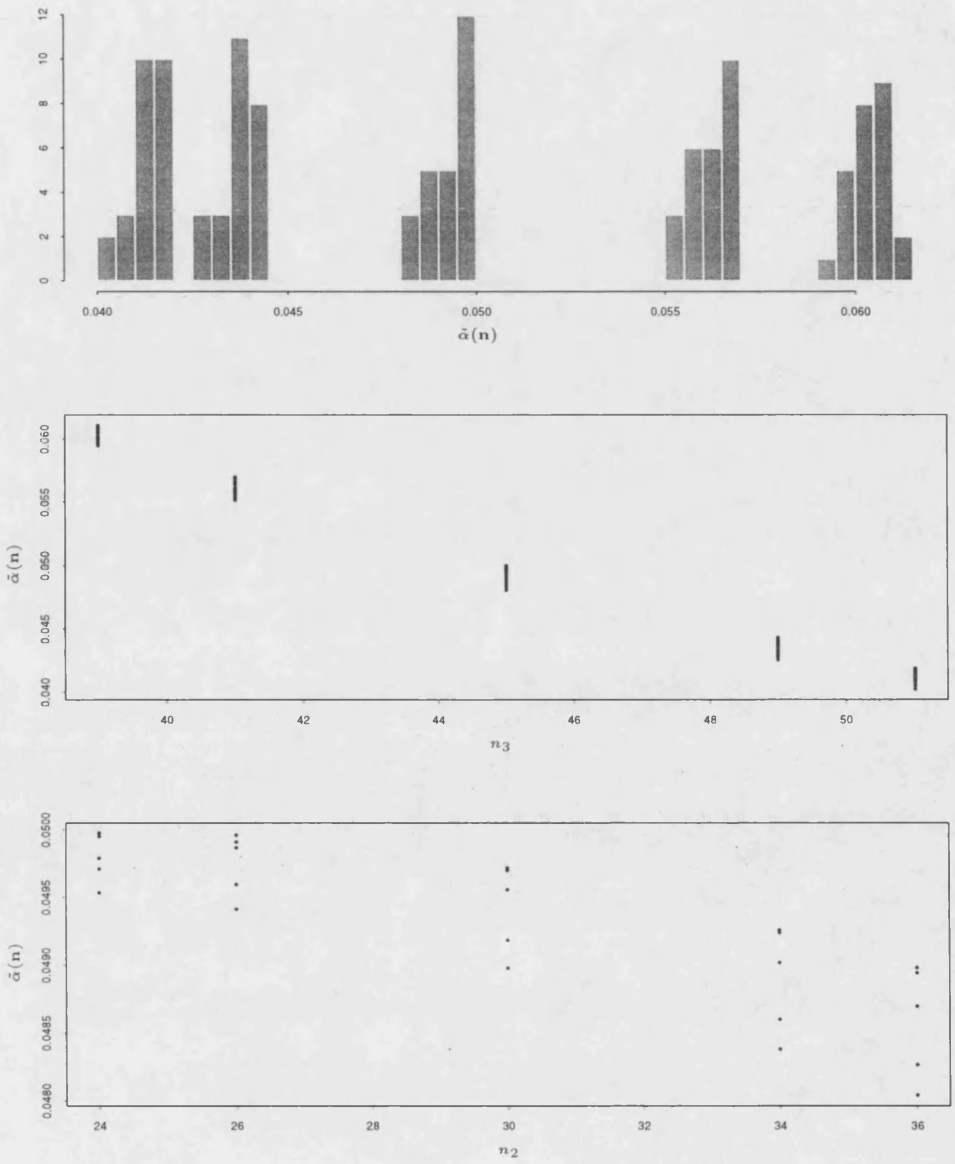


Figure 5-2: *Achieved type I error probability conditional upon sample size sequences for an optimal random group sequential test. The test sample sizes follow model 1 and the test is optimised for* F_1. *Overall type I and type II error probabilities are 0.05. The histogram and upper scatter plot show all values of* $\tilde{\alpha}(\mathbf{n})$, *while the lower graph shows only error probabilities for sample size sequences where* $n_3 = 45$.

test is symmetric, $\widetilde{\alpha}(\mathbf{n}) = \widetilde{\beta}(\mathbf{n})$ for any $\mathbf{n} = (n_{1,j_1}, n_{2,j_2}, n_{3,j_3})$. The histogram shows that achieved conditional error probabilities are sharply divided into discrete subsets, with the overall range of values of $\widetilde{\alpha}(\mathbf{n})$ being from 0.0402 to 0.0611. The subsets of conditional error probabilities correspond to differing maximum sample sizes, as shown in the centre graph in figure 5-2, which plots the achieved value of $\widetilde{\alpha}(\mathbf{n})$ against the maximum sample size n_3 of the sample size sequence $\mathbf{n}$. Larger maximum sample sizes result in lower conditional error probabilities, as would be expected. The lower graph in figure 5-2 is of achieved error probabilities against the value of n_2 for those sample size sequences which have $n_3 = 45$; note that the vertical scale of this plot is much larger than that for the graph above. This plot shows that the variability in error probability among sample size paths with the same maximum sample size is systematically affected by the sample size at the penultimate analysis, with lower $\widetilde{\alpha}(\mathbf{n})$ generally being associated with sample size sequences which have larger n_2. The same pattern is shown by plotting $\widetilde{\alpha}(\mathbf{n})$ against n_2 for other values of n_3.

Figure 5-3 shows the achieved values of $\widetilde{\alpha}(\mathbf{n})$ for random group sequential tests with sample sizes following models 2 and 3. Broadly similar patterns to those seen for model 1 in figure 5-2 are clearly visible, but the greater degree of uncertainty as to the possible maximum sample size results in the range of values of $\widetilde{\alpha}(\mathbf{n})$ being much greater. There is also far less distinction between the subsets of $\widetilde{\alpha}(\mathbf{n})$ values in the results for model 2, due to the fact that for any particular value of n_3, there are far more possible values of n_1 and n_2. In contrast, a particular value of n_3 in model 3 can only be achieved after a relatively small number of values of N_1 and N_2; in this case, the distinction between subsets of achieved $\widetilde{\alpha}(\mathbf{n})$ values is much greater. Similar patterns have been observed for models 4, 5, and 6, although the patterns are less distinct in tests following the five-analysis models. This is due to the fact that these models have a greater range of different possible sample size paths. In each case, the discretisation of $\widetilde{\alpha}(\mathbf{n})$ values was primarily caused by the achieved value of the maximum possible sample size n_K and was less clear for the larger values of the maximum sample size.

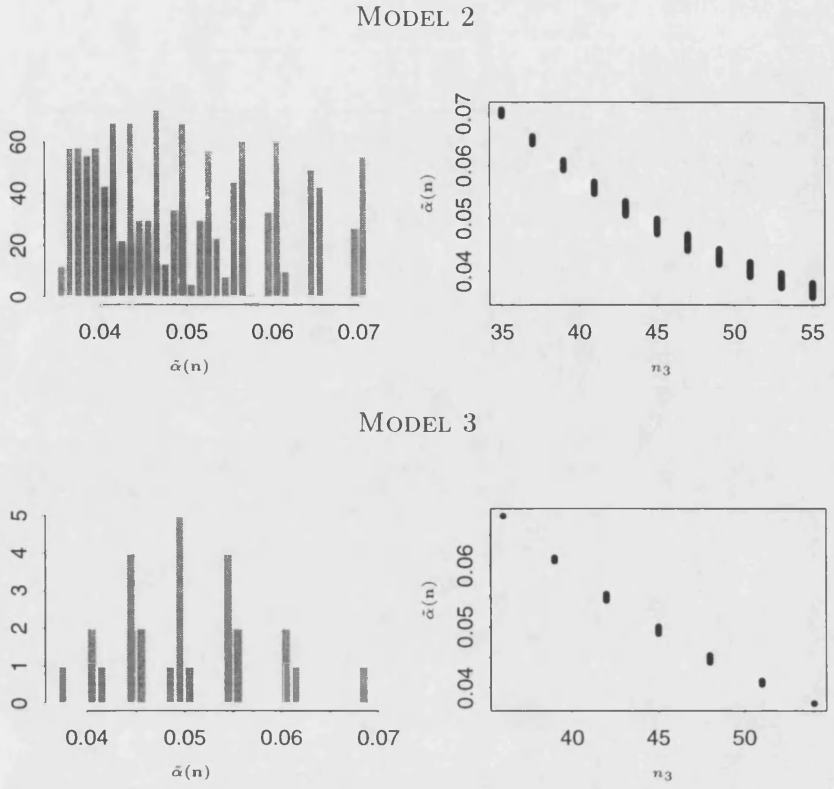


Figure 5-3: *Achieved type I error probability conditional upon sample size sequences for optimal random group sequential tests. The upper graphs are for a test with sample sizes following model 2, while the lower graphs are for a test following model 3. Both tests are optimised for* F_1. *Overall type I and type II error probabilities are 0.05.*

Random group sequential tests following sample size models 1 to 6 and minimising objective functions F_{21}, F_{31} and F_4 were also found. Recall that these objective functions are defined to be

$$F_{21} = \tfrac{1}{2}\left(\mathbb{E}_0\{N\} + \mathbb{E}_\delta\{N\}\right) \qquad F_{31} = \tfrac{1}{2}\left(\mathbb{E}_{-\delta/2}\{N\} + \mathbb{E}_{3\delta/2}\{N\}\right)$$

$$F_4 = \tfrac{1}{5}\mathbb{E}_{\delta/2}\{N\} + \tfrac{1}{10}\sum_{i=-2, i\neq 2}^{6} \mathbb{E}_{i\delta/4}\{N\}.$$

Since results for F_5 in chapters 3 and 4 were close to those for F_4, we have not

Model	Range of $\widetilde{\alpha}(\mathbf{n})$		Maximum sample size	
	F_1	F_{31}	n_K	n_K/n_{fix}
1	0.040 – 0.061	0.039 – 0.063	39 – 51	0.90 – 1.18
2	0.035 – 0.071	0.033 – 0.073	35 – 55	0.81 – 1.27
3	0.037 – 0.068	0.036 – 0.069	36 – 54	0.83 – 1.25
4	0.043 – 0.057	0.039 – 0.062	45 – 55	1.04 – 1.27
5	0.040 – 0.062	0.034 – 0.070	42 – 58	0.97 – 1.34
6	0.031 – 0.086	0.030 – 0.087	30 – 60	0.69 – 1.39

Table 5.1: *Tabulated values are maximum and minimum values of* $\widetilde{\alpha}(\mathbf{n})$ *and of* n_K *for random group sequential tests minimising* F_1 *and* F_{31}. *Three analysis tests considered follow sample size models one, two, and three, while five analysis tests follow sample size models four,five, and six.*

considered random group sequential tests optimised for F_5. As we only present results for symmetric tests, with $\alpha = \beta$, we do not consider the asymmetric objective functions F_{22}, F_{23}, F_{32}, and F_{33}.

The greatest range of values of $\widetilde{\alpha}(\mathbf{n})$ for the random group sequential tests was for tests minimising F_{31}, while the smallest range was for the tests minimising F_1. These values are shown in table 5.1 for tests following all the sample size models we defined in §5.1.5, along with the maximum sample sizes for each model. It is clear that the range of achieved $\widetilde{\alpha}(\mathbf{n})$ values is greater for those models with more uncertainty as to the maximum sample size which may be achieved by the test. Thus, in practice we would wish to control the maximum sample sizes possible under our sample size model in order to ensure the variation in achievable $\widetilde{\alpha}(\mathbf{n})$ values would be acceptable in frequentist terms.

5.2.2 Achieved objective function values

Following the notation used for the achieved error probability conditional on a particular sample size sequence $\mathbf{n} = (n_{1,j_1}, \ldots, n_{K,j_K})$, we define $\widetilde{F}_r(\mathbf{n})$ to be the value of objective function F_r achieved by the optimal random group sequential test when

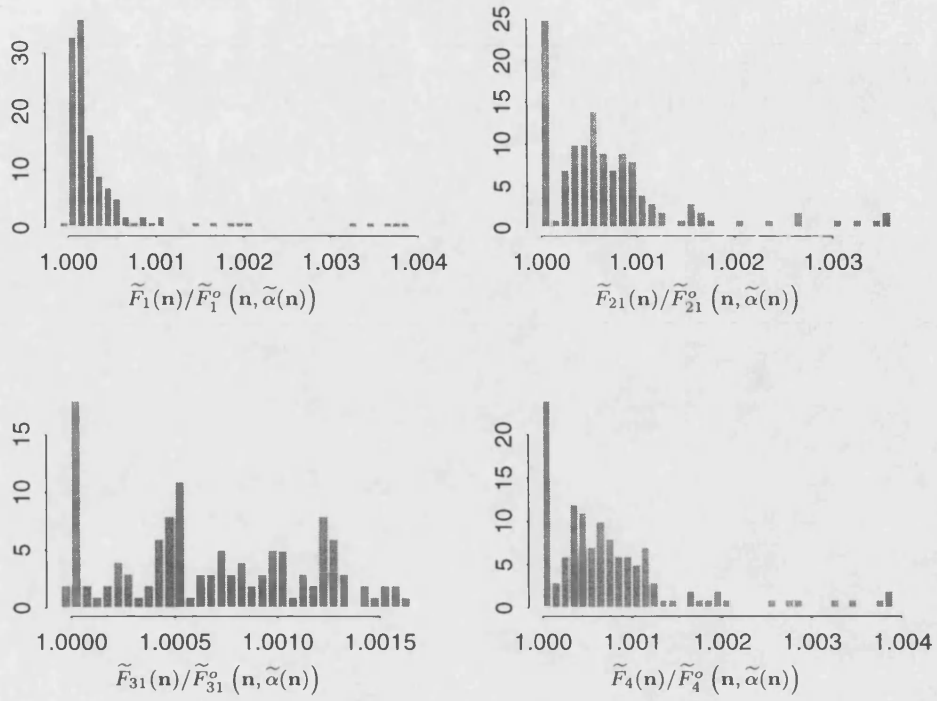


Figure 5-4: *Expected sample size results for the random group size test based on model 1 and optimising objective functions* F_1, F_{21}, F_{31} *and* F_4. *Overall error probabilities are* $\alpha = \beta = 0.05$. *Displayed values are ratios of* $\widetilde{F}_k(\mathbf{n})$ *to* $\widetilde{F}_k^o\left(\mathbf{n},\widetilde{\alpha}(\mathbf{n})\right)$ *for the objective functions considered.*

sample size sequence $\mathbf{n}$ is observed. We also define $\widetilde{F}_r^o\left(\mathbf{n},\widetilde{\alpha}(\mathbf{n})\right)$ to be the optimal value of F_r amongst all fixed group sequential tests with sample size sequence $\mathbf{n}$ and error probabilities $\widetilde{\alpha}(\mathbf{n})$ and $\widetilde{\beta}(\mathbf{n})$. That is, if we had known in advance what $\mathbf{n}$ would be, and had allowed the error probabilities $\widetilde{\alpha}(\mathbf{n})$ and $\widetilde{\beta}(\mathbf{n})$, $\widetilde{F}_r^o\left(\mathbf{n},\widetilde{\alpha}(\mathbf{n})\right)$ would be the minimum value of objective function F_r achievable. Finally, we shall denote the average values of $\widetilde{F}_r(\mathbf{n})$ and $\widetilde{F}_r^o\left(\mathbf{n},\widetilde{\alpha}(\mathbf{n})\right)$ by $\widetilde{F}_r$ and $\widetilde{F}_r^o$ respectively, the averaging being weighted by the sample size model chosen. We consider objective functions F_1, F_{21}, F_{31} and F_4. We use the superscript $\widetilde{\ }$ to indicate an achieved value of the objective function F_r, while the value of F_r achieved conditional upon sample size sequence $\mathbf{n}$ is indicated by the inclusion of $\mathbf{n}$.

Model	r	$\widetilde{F}_r$	$\widetilde{F}_r^o$	$\widetilde{F}_r/n_{fix}$
1	1	38.17	38.16	0.882
	21	31.64	31.62	0.731
	31	22.08	22.07	0.510
	4	31.08	31.07	0.718
2	1	38.47	38.46	0.889
	21	32.07	32.04	0.741
	31	22.70	22.68	0.524
	4	31.53	31.51	0.728
3	1	37.85	37.85	0.874
	21	30.97	30.97	0.715
	31	21.12	21.12	0.488
	4	30.41	30.41	0.702

Model	r	$\widetilde{F}_r$	$\widetilde{F}_r^o$	$\widetilde{F}_r/n_{fix}$
4	1	34.70	34.69	0.802
	21	26.92	26.89	0.622
	31	17.22	17.20	0.398
	4	26.79	26.77	0.619
5	1	34.72	34.71	0.802
	21	26.95	26.92	0.622
	31	17.31	17.29	0.400
	4	26.83	26.81	0.620
6	1	36.97	36.96	0.854
	21	29.62	29.60	0.684
	31	19.37	19.35	0.447
	4	29.16	29.15	0.674

Table 5.2: *Achieved objective function values for tests following sample size models 1 to 6 and optimised for objective functions* F_1, F_{21}, F_{31} *and* F_4. *Values are averaged over the sample size model and all tests have average error rates* $\alpha = \beta = 0.05$.

Figure 5-4 shows results for objective functions F_1, F_{21}, F_{31} and F_4 when the sample sizes follow model 1. The values given are ratios of the achieved $\widetilde{F}_r(\mathbf{n})$ to $\widetilde{F}_r^o(\mathbf{n}, \widetilde{\alpha}(\mathbf{n}))$ when $r = 1, 21, 31$ or 4. Clearly, the random group size test is near-optimal for each possible sample size sequence, with the achieved objective function value being within 0.4% of the optimal value in all cases.

Table 5.2 shows the average objective function values achieved by the optimal random group sequential tests averaged over all the sample size models considered. In each case, the average of the conditionally optimal objective function values, $\widetilde{F}_r^o$, is only slightly smaller than the achieved average, $\widetilde{F}_r$. It is clear that knowledge of $\mathbf{n}$ in $\widetilde{F}_r^o$ does not lead to a significant decrease in the average achieved objective function value.

5.2.3 Bayes risk of the optimal random group sequential tests

In the previous section, we considered the efficiency of the optimal random group sequential tests from the frequentist point of view, both conditionally upon the achieved

sample size sequence $\mathbf{n}$ and averaged over the sample size model. This efficiency was measured in term of the expected sample sizes of tests with matching error rates. We now consider an alternative efficiency measure, which combines the expected sample sizes and error rates of the tests.

Recall that we wish to decide between the decisions D_0: '$\mu = 0$' and D_δ: '$\mu = \delta$', with a prior $\pi(\mu)$ on μ. The costs of incorrectly deciding D_δ and D_0 are d_0 and d_δ respectively. Define the cost of wrong decision vector $\mathbf{d} = (d_0, d_\delta)$, and let $\mathbf{d}_{opt}$ be the value of $\mathbf{d}$ which gives the Bayes rule corresponding to our optimal random group sequential test. That is, $\mathbf{d}_{opt}$ is the decision cost vector which gives a Bayes rule with the probabilities of making a wrong decision α and β. We then write $\mathcal{R}_r^{(R)}(\mathbf{d}_{opt})$ for the Bayes risk of the decision rule which corresponds to our optimal random group sequential test optimising objective function F_r. The risk of a decision rule is simply the expected cost of the rule, and our problem has been constructed to have

$$\mathcal{R}_r^{(R)}(\mathbf{d}_{opt}) = k_1 \widetilde{F}_r + k_2 \left(d_\delta \pi(\delta) \mathbb{P}_\delta\{D_0\} + d_0 \pi(0) \mathbb{P}_0\{d_\delta\} \right), \tag{5.5}$$

as discussed in §5.1.2. The Bayes decision theory problem can be designed so that its solution is a decision rule which minimises any one of the objective functions we have considered. In equation (5.5) the achieved value of the objective function of interest is $\widetilde{F}_r$ and the values of the constants k_1 and k_2 depend upon which objective function we wish to minimise, as discussed on page 78. We also define the risk of this rule, conditional upon an observed sample size sequence $\mathbf{n}$ to be

$$\mathcal{R}_r^{(R)}(\mathbf{n}, \mathbf{d}_{opt}) = k_1 \widetilde{F}_r(\mathbf{n}) + k_2 \left(d_\delta \pi(\delta) \mathbb{P}_\delta\{D_0|\mathbf{n}\} + d_0 \pi(0) \mathbb{P}_0\{d_\delta|\mathbf{n}\} \right). \tag{5.6}$$

If we knew in advance what sample size sequence was to occur in the trial, the same decision cost vector would give a different Bayes rule to that obtained as a special case of the random groups Bayes rule. We can easily calculate this rule using the

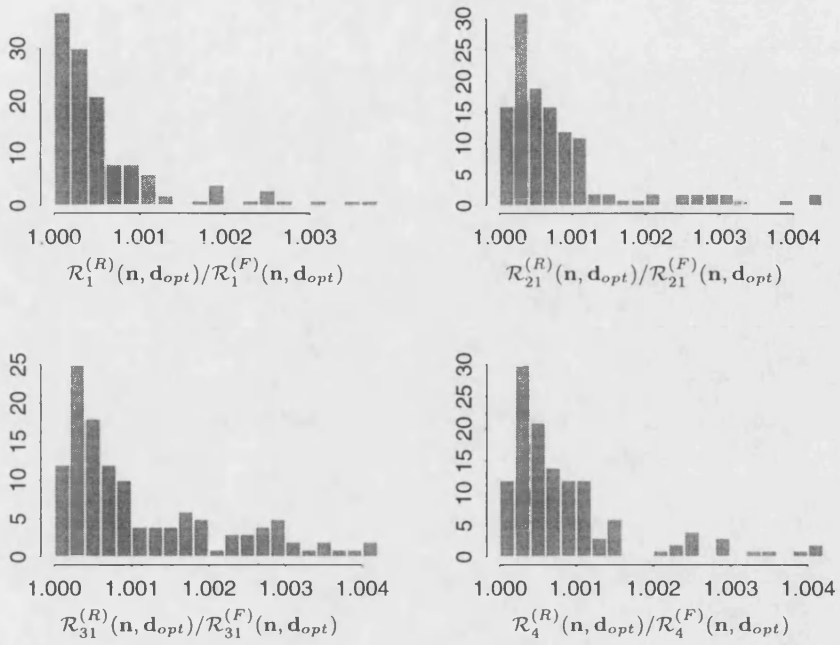


Figure 5-5: *Results for the Bayes rules minimising objective functions* F_1, F_{21}, F_{31}, *and* F_4. *Displayed values are* $\mathcal{R}_r^{(R)}(\mathbf{n}, \mathbf{d}_{opt})$ / $\mathcal{R}_r^{(F)}(\mathbf{n}, \mathbf{d}_{opt})$. *All risks are calculated using the decision cost vector* $\mathbf{d}_{opt}$, *which gives a Bayes rule corresponding to the optimal random group sequential test with error rates* $\alpha = \beta = 0.05$ *and following sample size model 1.*

backward induction algorithm from §5.1.3, optimising the objective function of our choice. Equation (5.6) can then be used to evaluate the risk of this Bayes rule, which we denote $\mathcal{R}_r^{(F)}(\mathbf{n}, \mathbf{d}_{opt})$.

Figure 5-5 shows the values of $\mathcal{R}_r^{(R)}(\mathbf{n}, \mathbf{d}_{opt})/\mathcal{R}_r^{(F)}(\mathbf{n}, \mathbf{d}_{opt})$ for the Bayes rule following sample size model 1 and optimising each of F_1, F_{21}, F_{31}, and F_4. These graphs bear a strong resemblance to those in figure 5-4. However, the values in figure 5-4 ratios of expected sample sizes for tests which had been forced to have matching error rates. The results in figure 5-5 are risk ratios between decision rules which do not have matching error rates, but have the same costs for wrong decisions. The risks take both expected sample sizes and the error probabilities into account. We are comparing the expected

costs of the overall Bayes procedure conditional upon a particular sample size sequence $\mathbf{n}$ to the optimal Bayes rule for that $\mathbf{n}$ given decision cost vector $\mathbf{d}_{opt}$. It is clear from these figures that the conditional risks $\mathcal{R}_r^{(R)}(\mathbf{n}, \mathbf{d}_{opt})$ are very close to the lowest possible risks for the same decision cost vector $\mathbf{d}_{opt}$ and sample size sequence $\mathbf{n}$. This indicates that the random groups Bayes rule defined by the decision cost vector $\mathbf{d}_{opt}$ is very close to optimal for each possible sample size sequence individually, as well as having the minimum risk when averaged over the sample size model. Similar patterns were observed for the other sample size models studied.

5.2.4 Conclusions

In §5.2.1 – §5.2.3, we have considered the performance of the optimal random group sequential tests conditional upon the sample size sequences observed, both from a frequentist and a Bayesian point of view. The achieved error probabilities conditional upon a specific sample size path $\mathbf{n}$ is mainly determined by the achieved value of the maximum sample size N_K; recall that due to the symmetry of our optimal random group sequential tests, $\tilde{\alpha}(\mathbf{n}) = \tilde{\beta}(\mathbf{n})$ for each sample size sequence $\mathbf{n}$. As the range of possible maximum sample sizes increases, so does the range of achieved error probabilities. In all cases, the achieved objective function values conditional upon a specific sample size path are very nearly optimal over all tests with the same sample size path and achieved error probabilities. In the Bayesian setting, we have seen that the risks conditional upon the observed sample size sequences are similarly close to optimality. The achieved objective function and risk values we have seen indicate that the optimal random group sequential tests are close to conditional optimality for each $\mathbf{n}$ both in terms of the expected sample size and expected cost. It is reassuring to see that these designs have both good Bayes and frequentist properties.

Part of our motivation in developing these optimal random group sequential tests was to provide a means of assessing the performance of existing methods when unanticipated

sample size sequences occur. However, the properties of the optimal random group sequential tests we have studied indicate that this method can be useful in its own right. As long as the range of possible maximum sample size values is controlled, the variation in achieved error rates will be within tolerable limits. Precisely how restricted the range of maximum sample sizes should be will depend on the degree of variation in achieved error probability that is considered acceptable.

Recall that our optimal random group sequential tests can be viewed as a collection of fixed group size tests. Each of these fixed group sequential tests will have expected sample size characteristics close to optimal among tests with the same sample size sequence $\mathbf{n}$ and error probabilities $\tilde{\alpha}(\mathbf{n})$. The optimal random group sequential test minimises our chosen objective function among all tests which have our target error rates when averaged over the sample size model. Thus, in frequentist terms, we have a test which is optimal when averaged over all the possible sample size sequences in our chosen model and near-optimal conditional upon $\mathbf{n}$ and $\tilde{\alpha}(\mathbf{n})$. In the Bayesian setting, we have seen that the risks of these procedures will also be near optimal for each sample size sequence.

5.3 Performance of existing methods

The error spending method of Lan & DeMets (1983), discussed in §2.3, is designed to deal with unanticipated sample size sequences. However, the Δ-family of Pampallona & Tsiatis (1994) discussed in §2.2 needs further modification to deal with an unanticipated sample size sequence. In §5.3.1, we discuss one method of adapting the Δ-family to unanticipated group sizes. We then go on to consider how well the Δ-family and error-spending tests preserve their nominal error rates when the group sizes deviate from the planned sequence in §5.3.2. Then, in §5.3.3, we compare the Δ-family and error spending tests to the optimal random group sequential tests defined in §5.1.

5.3.1 Adapting the Δ-family tests to random group sizes

The Δ-family of tests, as described by Pampallona & Tsiatis (1994), is defined for a sequence of equally sized groups of observations. As we noted in §2.2, the method can easily be generalised to include non-equal group sizes, if the schedule of analyses is known in advance, but further modification is required to adapt the method to an observed sequence of sample sizes which differs from the planned sequence. In this section, we discuss a method of adapting the Δ-family tests to unanticipated group sizes from Jennison & Turnbull (2000, chapter 4).

We wish to test the null hypothesis H_0:$\mu \leq 0$ against the alternative hypothesis H_1:$\mu > 0$ with type I and II error probabilities α and β at $\mu = 0$ and $\mu = \delta$ respectively. We can take observations $X_1, X_2, \ldots$, with each observation being independent and identically $N(\mu, \sigma^2)$ distributed, where we assume that σ^2 is known. We define the summary statistic S_n to be the sum of the first n observations and n_{fix} to be the number of observations required for the equivalent non-sequential design. Recall from the description of the Δ-family in §2.2 that by choosing a value of the parameter Δ, we fix n_K and define the boundary points at analysis i from the equations

$$a_i = \delta n_i - c_2 i^{\Delta - 1/2}\sqrt{n_i \sigma^2}, \qquad \text{and} \qquad b_i = c_1 i^{\Delta - 1/2}\sqrt{n_i \sigma^2},$$

where constants c_1 and c_2 are determined by the choice of Δ, as discussed in §2.2.

In designing a trial, we plan K analyses at sample sizes $\mathbf{n} = (n_1, n_2, \ldots, n_K)$, with $n_1 < n_{fix}$ and $n_K > n_{fix}$. We can find a value $\Delta(\mathbf{n})$ of Δ to give a test which has interim analyses at the desired sample sizes and achieves the desired error probabilities. This gives rise to a group sequential boundary described by the boundary points $\{(a_i, b_i); i = 1, \ldots, K\}$. We then define the values $\zeta_i^{(u)}$ and $\zeta_i^{(l)}$ to be

$$\zeta_i^{(u)} = \mathbb{P}_0\{S_{n_i} > b_i\} \qquad \text{and} \qquad \zeta_i^{(l)} = \mathbb{P}_0\{S_{n_i} < a_i\},$$

noting that these probabilities are calculated using the marginal distribution of S_{n_i} without considering the possibility of early stopping rather than the joint distribution of $\{S_{n_1}, \ldots, S_{n_K}\}$.

During the trial, instead of taking observations in groups according to the planned schedule of analyses, we may instead see n'_1 observations at the first analysis and so on, resulting in an achieved sample size path $\mathbf{n}' = (n'_1, \ldots, n'_K)$. At analysis i, we modify our boundary to allow for the achieved sample size by using boundary points a'_i, b'_i, defined by the equations

$$\mathbb{P}_0\left\{S_{n'_i} > b'_i\right\} = \zeta_i^{(u)} \qquad \text{and} \qquad \mathbb{P}_0\left\{S_{n'_i} < a'_i\right\} = \zeta_i^{(l)}.$$

Detailed formulae for the boundary points are given by Jennison & Turnbull (Jennison & Turnbull, 2000, p. 96).

With this approach, we are keeping the boundary points constant on a standardised scale. Since both upper and lower boundaries are standardised under $\mu = 0$, the boundaries will still converge at the final analysis. This approach also ensures that the type I error probability is kept close to the nominal value α, as we shall see in §5.3.2. This "significance level" method was been used by Pocock (1977) to modify group sequential tests to allow for analyses which do not comply with the planned schedule of group sizes. A different approach was taken by Emerson & Fleming (1989). These authors interpolated between boundary points of a repeated significance test scheme to cope with unpredictable group sizes, but this approach was less effective at preserving type I and type II error than the significance level approach.

5.3.2 Deviations from nominal error rates for existing methods

The error spending method was designed to deal with trials where the observed sequence of sample sizes differs from that planned, and we have described a means of adapting

a Δ-family test to an unplanned sequence of sample sizes in §5.3.1. It is important to consider how well these methods cope when the sample size sequence departs from the planned schedule of analyses. In this section, we consider how well the error spending and modified Δ-family tests preserve the intended type I and type II error probabilities.

Usually, the primary concern is for the preservation of the type I error probability at or near its nominal level. The error spending method is designed to preserve type I error probability at exactly the nominal level regardless of the sample size sequence observed, but the type II error probability may vary freely. The method described in the previous section for adapting a Δ-family test to unexpected sample size sequences is approximate and hence does not preserve the type I error exactly. As we shall see, however, the modified Δ-family does well at preserving the type I error close to the intended value α.

To apply the Δ-family to one of the sample size models described in §5.1.5, we find the value of Δ which will give us a test with the required type I and type II error probabilities if the most likely sample size sequence is observed. Once this value of Δ is found, we can adapt the resulting test to the observed sample size sequence as described in §5.3.1. In all the models we have considered, this most likely sequence consists of equally sized groups of observations. For models 1, 2, and 3 the most likely sample size sequence is $\mathbf{n} = (15, 30, 45)$, for models 4 and 5 the most likely sample size sequence is $\mathbf{n} = (10, 20, 30, 40, 50)$, and for model six $\mathbf{n} = (9, 18, 27, 36, 45)$ is sequence most likely to occur. A value of $\Delta = -0.07$ specifies the desired test for models 1, 2, and 3, $\Delta = 0.18$ results in the desired test for models 4 and 5 and for model 6 the desired test has $\Delta = -0.25$.

The histograms in figure 5-6 show the range of achieved type I and type II error probabilities of the Δ-family tests when applied to models 2, 4 and 6 with equal nominal type I and type II error probabilities $\alpha = \beta = 0.05$. Recall from §5.2 that we define $\widetilde{\alpha}(\mathbf{n})$ and $\widetilde{\beta}(\mathbf{n})$ to be the error probabilities achieved by the test when sample size

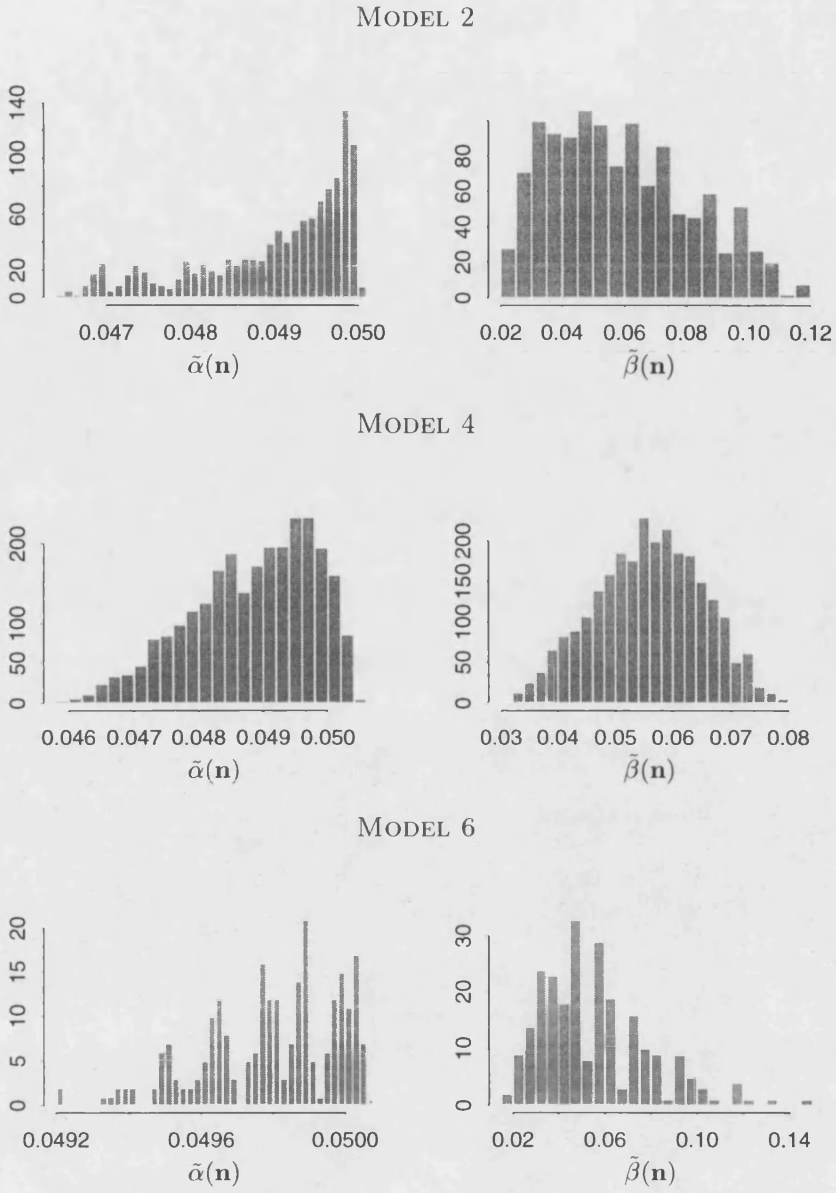


Figure 5-6: *Achieved values of* $\tilde{\alpha}(\mathbf{n})$ *and* $\tilde{\beta}(\mathbf{n})$ *for* Δ*-family tests not following the planned sample size sequence. The top histograms includes all sample size sequences in model 2, while the middle and lower histograms include all sample size sequences in models 4 and 6 respectively. All tests were designed for the most likely sequence of sample sizes for the respective models, with nominal error probabilities* $\alpha = \beta = 0.05$.

sequence $\mathbf{n}$ is observed. In each case, the range of observed values of $\widetilde{\alpha}(\mathbf{n})$ is quite small. For model 2, $\widetilde{\alpha}(\mathbf{n})$ lies between 0.046 and 0.050, for model 4 $\widetilde{\alpha}(\mathbf{n})$ lies between 0.046 and 0.050 and for model 6 $\widetilde{\alpha}(\mathbf{n})$ lies between 0.049 and 0.050. Thus, we can see that even in the worst case the achieved type I error probability is within 10% of its nominal value and that the deviations seen are almost always conservative. In cases where the observed ratios $n_i : n_{i+1}$ $(i = 1, \ldots, K-1)$ are the same as in the planned sample size sequence, $\widetilde{\alpha}(\mathbf{n}) = \alpha$, that is that the type I error probability is preserved at exactly the nominal level. The deviation from the nominal value of α increases as the ratios between the observed group sizes deviate further from the planned schedule.

However, figure 5-6 also shows that the type II error probabilities achieved by the Δ-family test vary much more than the type I error probabilities. The ranges of $\widetilde{\beta}(\mathbf{n})$ are 0.026 — 0.120 for model 2, 0.032 — 0.082 for model 4 and 0.016 — 0.150 for model 6. Unsurprisingly, the achieved value of $\widetilde{\beta}(\mathbf{n})$ is very strongly influenced by the maximum sample size n_3, with lower type II error probability occurring when the maximum sample size is greater than initially planned. Similar patterns of deviation in $\widetilde{\alpha}(\mathbf{n})$ and $\widetilde{\beta}(\mathbf{n})$ are seen for the other models considered. In general, models with a greater range of different sample size sequences result in a larger range of observed conditional error rates.

The error spending method is designed to preserve type I error probability exactly, regardless of the observed sample size sequence. However, the type II error probability can vary freely with the number of observations seen. In order to assess the variability of the achieved type II error probabilities under different sample size sequences, for each of models 1 to 6, error spending tests using both the γ-family and ρ-family of error spending functions were defined in the same way as the Δ-family tests. In each case a value of the relevant design parameter was chosen to specify a test which had the nominal error probabilities under the most likely sample size sequence. The values of γ which specified the desired tests were $\gamma = -3.23$ for models 1, 2, and 3, $\gamma = -1.13$

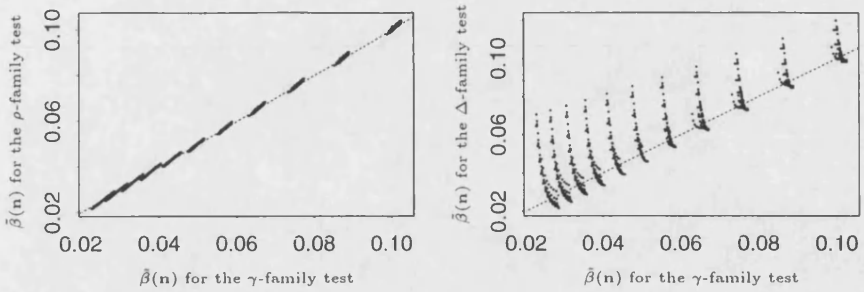


Figure 5-7: *Achieved values of $\widetilde{\beta}(\mathbf{n})$ for the Δ-family and error spending tests. Tests were designed for three analyses, each group consisting of 15 observations, with type I and type II errors $\alpha = \beta = 0.05$. Plotted values are of $\widetilde{\beta}(\mathbf{n})$ calculated for each possible sample size in model 2. The dotted lines indicate points where the tests have equal type II error.*

for models 4 and 5 and $\gamma = -3.99$ for model 6. The equivalent values of ρ were 2.71, 1.48, and 3.35 respectively. These tests were then applied to the cases where the other possible sample size sequences occurred. In each case, the range of possible values of $\widetilde{\beta}(\mathbf{n})$ for both γ-family and ρ-family error spending tests was less than for the Δ-family tests, despite the error spending tests having no flexibility as to the achieved type I error probability. The γ-family tests had a smaller range of values of $\widetilde{\beta}(\mathbf{n})$ than the ρ-family for models 1, 2, and 4 and a slightly larger range of $\widetilde{\beta}(\mathbf{n})$ values for models 3, 5, and 6.

For each of the sample size models considered, the tests using the γ-family and ρ-family error spending functions achieved very similar values of $\widetilde{\beta}(\mathbf{n})$ for any specific sample size sequence $\mathbf{n}$. This is demonstrated in the left hand plot in figure 5-7, which shows the values of $\widetilde{\beta}(\mathbf{n})$ achieved by error spending tests using both γ-family and ρ-family error spending functions designed for sample size sequence $\mathbf{n} = (15, 30, 45)$ and applied to the possible sample size sequences found in model 1. The achieved value of $\widetilde{\beta}(\mathbf{n})$ was also much more closely linked to the maximum sample size than was the case for

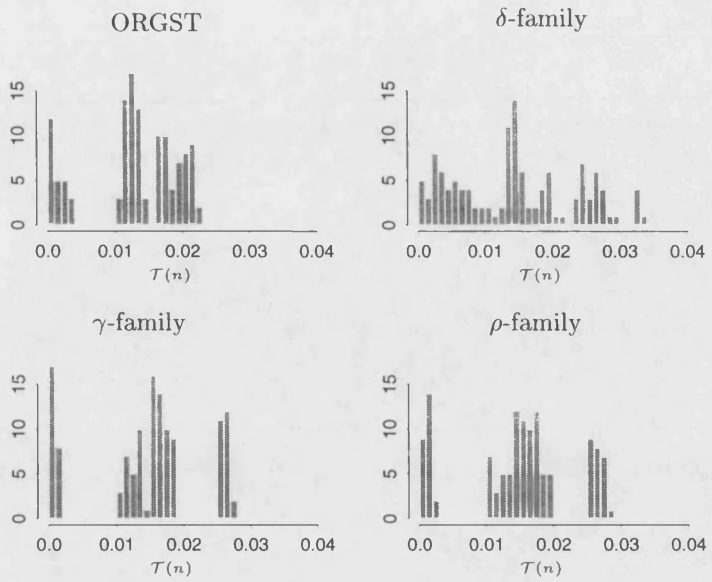


Figure 5-8: *Total deviation from the nominal error probabilities for tests based on sample size model one. Tests considered are the optimal random groups sequential test optimised for F_1 (denoted ORGST) and the Δ-family, γ-family and ρ-family tests designed for three groups of 15 observations each. The optimal test was designed with overall error rates $\alpha = \beta = 0.05$, while the other tests had these error rates under the most likely sample size sequence.*

the Δ-family tests and was generally lower for the error spending tests than for the Δ-family tests, as is shown in the right hand side of figure 5-7. This plots the value of $\widetilde{\beta}(\mathbf{n})$ achieved by the Δ-family test against those achieved by the γ-family error spending test under the same circumstances as in the left hand plot.

The error spending tests have less variation in $\widetilde{\beta}(\mathbf{n})$ values than the Δ-family tests as the error spending method responds to smaller than anticipated groups of observations by "spending" less type I and type II error probability than planned. The modified Δ-family test still uses the same significance levels if a smaller group of observations is seen, which is a less efficient use of the information contained within the data.

In order to assess the overall degree of variability in error probabilities we define

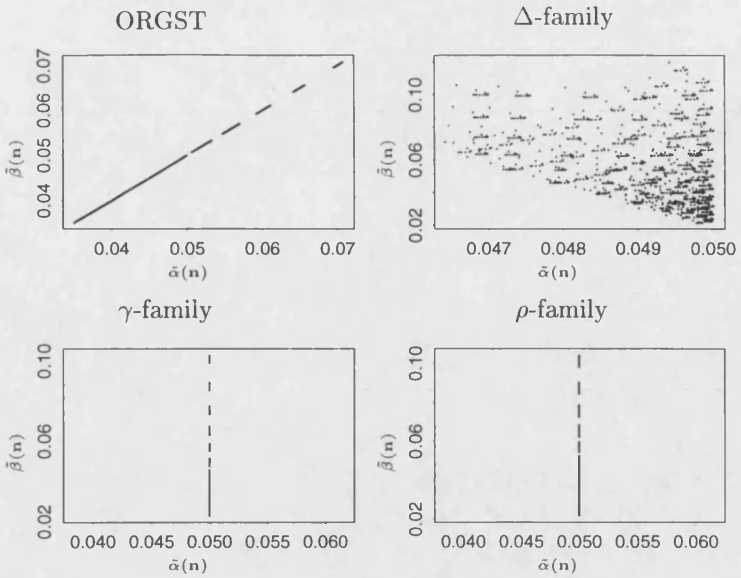


Figure 5-9: *Achieved* $\widetilde{\alpha}(\mathbf{n})$ *and* $\widetilde{\beta}(\mathbf{n})$ *for optimal random group sequential tests,* Δ*-family tests and error spending tests. The optimal random group sequential test was designed for model 2 with overall error rates* $\alpha = \beta = 0.05$*. The* Δ*-family and error spending tests were designed for three groups of 15 observations each, with* $\alpha = \beta = 0.05$*, and evaluated for all sample size sequences in model 2.*

the total deviation from the nominal error probability for any specific $\mathbf{n}$ to be $\mathcal{T}(\mathbf{n}) = |\widetilde{\alpha}(\mathbf{n}) - \alpha| + |\widetilde{\beta}(\mathbf{n}) - \beta|$. Figure 5-8 shows this value calculated for all the possible sample size sequences in model 1, evaluated for the optimal random group sequential test optimised for objective functions F_1 and for the Δ-family, γ-family and ρ-family tests. The smallest variability from the nominal error probability is seen for the optimal random group sequential test, with the greatest variation seen for the Δ-family tests. Of the error spending tests, there is slightly more variability for the ρ-family test. The total deviation from the nominal error probabilities for the other optimal random group sequential tests was also lower than for the error spending methods, although there was little difference between the random group sequential test optimised for F_{31} and the γ-family test. The same pattern was observed for the other sample size models.

Figure 5-9 compares the variability of achieved error rates for the optimal random group sequential tests, the Δ-family tests and the error spending tests. The achieved type I error is constant or nearly so for the Δ-family and error spending tests, while for the optimal random group sequential tests both error rates vary, with $\widetilde{\alpha}(\mathbf{n})= \widetilde{\beta}(\mathbf{n})$. This means that the optimal tests have less variation in $\widetilde{\beta}(\mathbf{n})$ than the other methods, but if precise control of type I error is desired, this may make the optimal tests inappropriate.

5.3.3 Comparing the efficiencies of existing methods and optimal tests

In chapters 3 and 4 and in §5.2.2 we have measured the efficiency of group sequential designs by considering the expected sample size averaged over values of μ. These average expected sample sizes were referred to as objective functions. We cannot compare the achieved objective function values for the optimal random group sequential tests directly to the values achieved by the Δ-family and error spending tests since each test has different average error probabilities. Instead, we turn to the Bayesian formulation of the problem and consider the risk or expected cost of each test. Recall that the risk of a decision rule is a weighted sum of the achieved objective function value and achieved error rates averaged over the sample size model, as given in equation (5.5) on page 100. For each combination of sample size model and objective function, there is a unique decision cost vector $\mathbf{d}_{opt}$ for which the Bayes decision rule minimising the chosen objective function when the observed sample size sequences follow the specified model has the desired probabilities α and β of making a wrong decision. All the risks in this section are calculated using the decision cost vector $\mathbf{d}_{opt}$ from the Bayes rule corresponding to the relevant optimal random group sequential test.

Table 5.3 shows the achieved risks for Δ-family, γ-family and ρ-family tests applied to sample size models 1 to 6, given as percentages of the risk of the relevant optimal random group sequential tests. In each case, the tests are symmetric, designed for overall type I and type II error probability 0.05, although for many sample size paths

Model		Δ	γ	ρ
1	F_1	101.4	100.7	100.8
	F_{21}	103.0	101.0	101.5
	F_{31}	106.7	102.4	103.9
	F_4	103.1	101.0	101.6
2	F_1	102.1	101.0	101.1
	F_{21}	103.6	101.4	101.9
	F_{31}	107.0	102.7	104.1
	F_4	103.7	101.4	101.9
3	F_1	101.0	100.6	100.6
	F_{21}	102.6	100.8	101.3
	F_{31}	106.8	102.1	103.5
	F_4	102.8	100.8	101.3
4	F_1	101.2	100.8	100.8
	F_{21}	103.0	101.1	101.4
	F_{31}	109.5	104.2	105.4
	F_4	102.8	100.9	101.2
5	F_1	101.2	100.8	100.8
	F_{21}	103.2	101.2	101.5
	F_{31}	109.7	104.3	105.5
	F_4	103.0	101.0	101.3
6	F_1	102.0	101.2	101.3
	F_{21}	104.2	101.7	102.3
	F_{31}	109.5	103.2	105.2
	F_4	104.4	101.6	102.3

Table 5.3: *Tabulated values are average risks of random group sequential tests based on the Δ-family, γ-family, and ρ-family designs, given as percentages of the risk of the relevant optimal random group sequential test. All risks are evaluated using the cost of a wrong decision from the appropriate optimal random group sequential test.*

the achieved values of $\widetilde{\alpha}(\mathbf{n})$ and $\widetilde{\beta}(\mathbf{n})$ will be different for the non-optimal tests. In each case the greatest risk is for the Δ-family tests, while the risks of the error spending tests are similar in all cases, and are noticeably lower than those for the Δ-family. Of the two families of error spending functions, the γ-family has slightly lower risks. While it is difficult to compare the results for the three- and five-analysis tests, it seems that overall the five-analysis tests have lower relative risks.

An obvious question is how much of the departure from optimality seen in table 5.3 is due to the Δ-family and error spending tests being non-optimal for the sample size sequences for which they were designed, and how much is due to the effect of unanticipated sample size sequences. To address this question, we consider the risks of the Δ-, γ-, and ρ-family tests when applied to the single sample size sequence for which they were designed. These risks are in table 5.4, given as percentages of the risks of the optimal fixed group sequential tests designed for the same sample size sequence and error rates. The risks are calculated using the decision cost vector $\mathbf{d}_{opt}$ which gives

Model		Δ	γ	ρ
1	F_1	100.4	100.0	100.0
	F_{21}	101.9	100.2	100.7
	F_{31}	106.0	101.3	102.7
	F_4	102.1	100.2	100.7
2	F_1	100.4	100.0	100.0
	F_{21}	101.9	100.2	100.6
	F_{31}	105.8	101.3	102.6
	F_4	102.1	100.2	100.7
3	F_1	100.4	100.0	100.0
	F_{21}	101.9	100.2	100.7
	F_{31}	106.3	101.4	102.8
	F_4	102.1	100.2	100.7

Model		Δ	γ	ρ
4	F_1	100.4	100.3	100.1
	F_{21}	102.6	100.5	100.8
	F_{31}	110.0	103.4	104.7
	F_4	102.4	100.3	100.6
5	F_1	100.4	100.3	100.1
	F_{21}	102.6	100.5	100.8
	F_{31}	110.1	103.4	104.7
	F_4	102.4	100.3	100.6
6	F_1	100.8	100.1	100.2
	F_{21}	102.8	100.5	101.1
	F_{31}	108.2	101.9	103.9
	F_4	103.1	100.4	101.1

Table 5.4: *Tabulated values are risks of fixed group sequential tests from the Δ-, γ-, and ρ-families, designed for the most likely sample size sequence in each model, given as percentages of the risk of the relevant optimal fixed group sequential test. All risks are evaluated using the cost of a wrong decision from the appropriate optimal random group sequential test.*

the relevant optimal random group sequential test in each case. These are the same decision costs which were used in calculating the results in table 5.3. When we consider the results relating to the minimisation of F_1, we can see that the bulk of the departure from optimality is due to the random sample sizes, although it is here that the risks are closest to the minimum possible values. In the case of objective functions F_{21} and F_4, the non-optimality seems to be roughly evenly attributable to both sources and in the case of F_{31} the majority of the departure from optimality seems to be caused by the non-optimal nature of the Δ-, γ-, and ρ-family tests in the fixed groups setting.

Figure 5-10 shows the conditional risks of the optimal random group sequential and γ-family tests conditional upon the sample size sequence $\mathbf{n}$. The tests in the left plot follow model 1, while those in the right plot follow model 6. The optimal test following model 1 minimises F_{21} and the optimal test following model 6 minimises F_1, both tests having average error probabilities $\alpha = \beta = 0.05$. The error spending tests have been designed to have these error rates for the most likely sequences of sample sizes in

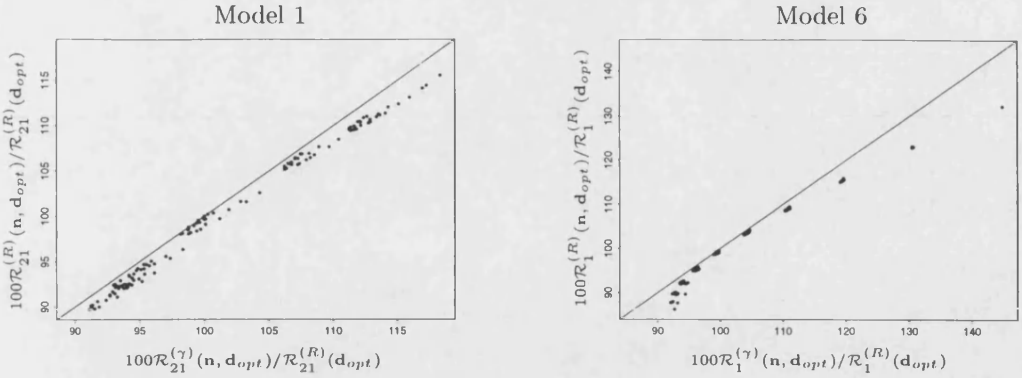


Figure 5-10: *Risks of the optimal random group sequential tests and γ-family error spending tests conditional upon the sample size sequence* $\mathbf{n}$. *Tests in the left hand plot follow model 1 and the optimal test here minimised* F_{21}. *In the right hand plot the optimal test minimises* F_1, *with both tests following sample size model 6. The optimal tests have average error probabilities* $\alpha = \beta = 0.05$. *Error spending tests have these error probabilities under the most likely sample size sequence from each model. All risks are given as percentages of the average risk of the relevant optimal random group sequential test, and are calculated using the decision costs of this optimal test.*

models 1 and 6, $\mathbf{n} = (15, 30, 45)$ and $\mathbf{n} = (9, 18, 27, 36, 45)$ respectively. All risks are calculated using the cost of wrong decision vector $\mathbf{d}_{opt}$ which gives rise to the relevant optimal random group sequential tests; these are the same costs as were used in the calculation of tables 5.3 and 5.4. Values are given as percentages of the average risk of this optimal test, $\mathcal{R}_{21}^{(R)}(\mathbf{d}_{opt})$ for the tests following model 1 and $\mathcal{R}_{1}^{(R)}(\mathbf{d}_{opt})$ for the tests following model 6.

The plots in figure 5-10 show that the conditional risks of the γ-family tests are greater than those of the optimal random group sequential tests for any sample size sequence. However, the discrepancy in risks is not large in any of the cases represented in these plots. In all the examples we have seen, the conditional risk of the γ-family test has never been more than approximately 4% greater than the conditional risk of the corresponding optimal group sequential test for the same sample size sequence.

As with other properties of random group sequential tests, the conditional risks are

strongly correlated with the maximum sample size. Both plots in figure 5-10 show this, with each cluster of points corresponding to different values of N_K. This is especially clear in the right hand plot. Larger risks occur when N_K is smaller; a larger value of N_K leads to more confidence in our decision and hence a lower risk, although there may be a higher expected sample size. However in chapter 3 the minimum value of F_1 we observed with $K = 5$ occurred when $n_K = 1.3 \times n_{fix}$. In this case, this corresponds to $n_K = 56.3$; of the models we consider only model 6 allows this many observations.

It is also clear from figure 5-10 that there is a greater discrepancy between the conditional risks of the γ-family and the optimal random group sequential tests when the maximum sample size is further from the most likely value. The error spending tests have been designed to reach this value and hence it is not surprising their performance deteriorates when we see greater or smaller maximum sample sizes. The degradation of the performance of the γ-family tests is greater when we see less than the expected number of observations than when we exceed this value.

5.3.4 Conclusions

Comparing the Δ-family tests to the optimal random group sequential tests we have found in this chapter, we can see that the Δ-family tests have the greatest range of deviation from the nominal error probabilities and the greatest increase in risk over the optimal random group sequential tests; the average risks for the Δ-family tests are up to 10% greater than the optimal values. It is not surprising that the Δ-family performs poorly in this setting, since the method was designed for a fixed sample size sequence. We have also considered applying the significance level approach which was used to adapt the Δ-family tests to the optimal fixed group sequential tests we found in chapter 3. The resulting random group sequential tests had lower risks than the Δ-family tests, but greater risks than the error spending tests. This is because, like the Δ-family tests, these adapted optimal fixed group sequential designs could not react to

larger or smaller groups of observations by altering the amount of type I error spent at each analysis, as the error spending method does.

The error spending method performs well, with the γ-family tests being slightly superior the ρ-family tests in most cases. Average risks for the γ-family within 5% of the risks of the optimal random group sequential tests, although this is at the cost of a greater variation in error probabilities than is seen for the optimal tests.

5.4 Discussion

We have found a means to minimise the expected sample sizes of group sequential tests when the number of observations seen at each analysis is not determined in advance. Our method does require a fixed maximum number of analyses, and that we can describe the possible total number of observations at each analysis by a simple probability distribution. Our goal in finding these optimal random group sequential tests was to provide a tool for assessing the efficiency of the Δ-family and error spending tests under these circumstances. However, we have seen that the overall error fluctuation of the optimal random group sequential tests is less than that for the Δ-family and error spending tests, albeit with a much greater variation in type I error. This variability is closely linked to the the range of maximum sample sizes possible under the sample size model. If the maximum total number of observations is not too variable, out optimal random group sequential designs would be viable procedures to use in monitoring a clinical trial, so long as it would be acceptable to have a test with, say, a $\pm 10\%$ fluctuation in type I error.

If the primary concern in the design of a group sequential clinical trial is the preservations of type I error at exactly the nominal level, then the error spending approach is to be preferred. Our results indicate that the γ-family tests have slightly superior properties than the ρ-family tests. While the Δ-family tests have been shown

to have less desirable properties than the optimal random group sequential designs and error spending tests, our results indicate that the significance level approach can be used to generalise a Δ-family tests to unanticipated group sizes with only a small to moderate impact upon their performance. This is reassuring for any researcher who has designed a trial using a Δ-family test but has then been faced with a departure from the planned schedule of analyses.

It is perhaps surprising that the random group sequential tests have so little variation in achieved error probabilities, since these tests can vary the conditional error rates in search of the test which minimises the relevant objective function value while maintaining the overall error probabilities. Given this, we might have expected to see greater variation in the values of $\tilde{\alpha}(\mathbf{n})$ for the optimal tests.

The sample size models considered in this chapter are relatively simple and much more sophisticated models could be considered. These models do not have to be restricted to integer sample sizes, so the method we have used could be extended to situations where the data are more complicated than the independent and identically distributed normal random variates we have considered. In such situations, we would measure groups in terms of information rather than the number of observations, as discussed on page 4. By considering models with a large number of possible "sample" sizes at each analysis, with little difference between them, we could approximate situations where the observed information statistic is a continuous function of the data.

The remaining advantage of the error spending approach is that the maximum number of analyses is not fixed, allowing as many analyses as are needed to meet a pre-specified information level. Our random group sequential tests are still restricted to a fixed maximum number of analyses. In the next chapter, we shall look at a method of extending our optimal tests to relax the requirement of a fixing a maximum number of analyses.

Chapter 6

A maximum information design

A popular approach for error spending test is the "maximum information" method of sampling (Lan & Zucker, 1993; Lan, Reboussin, & DeMets, 1994). In such a trial, sampling continues until a pre-specified level of information has been reached. In the simple case of independent and identically distributed normal data, this happens when the cumulative number of observations seen reaches a value fixed as part of the trial design. A characteristic of these maximum information designs is that the maximum number of observations K is variable, rather than being fixed. This gives a very flexible sequential design and it is intuitively appealing to take this approach. In order to examine the properties of the error spending tests, we shall develop optimal tests which follow this sampling scheme and accrue observations until a threshold value n_T is reached. The cumulative sample size at analysis i is denoted n_i, thus we observe a non-decreasing sequence of cumulative sample sizes $n_1, n_2, \ldots$. We shall terminate the trial at the first analysis i where $n_i \geq n_T$, if the trial has not already terminated due to crossing the sequential boundary. Throughout the remainder of this thesis we refer to these designs as threshold group sequential tests; we find that the term "maximum information tests" can be misleading as the pre-specified "maximum" information level may be exceeded.

In §6.1 we look at the appropriate Bayesian decision theory problem, the backward induction algorithm used to solve it and some suitable group size models. Then, in §6.2, we look at the overall and conditional properties of the resulting optimal tests and in §6.3 we compare our optimal tests to the error spending tests using the γ-family and ρ-family error spending functions. We no longer consider the Δ-family tests, as they are dependent upon a fixed maximum number of analyses.

6.1 Optimising to an information threshold

We now consider a method for designing an optimal group sequential test where the maximum number of analyses is not fixed as part of the test design. Instead, we specify a threshold of observations and terminate the trial at the first analysis where this number of observations is exceeded, if the trial has not already stopped due to crossing the group sequential boundary. In chapter 5, we specified models to describe the possible sequences of sample sizes which might be observed in the course of the trial. We shall use the same idea in this chapter, however the models we shall consider are more conveniently expressed in terms of the possible group size sequences. In §6.1.1, we define the frequentist problem we wish to solve and in §6.1.2 we state the Bayesian decision theory problem which has the same solution as our frequentist problem. The backward induction algorithm we use to find the Bayes rule of interest is described in §6.1.3, while in §6.1.4 we define several group size models which we shall use to explore the properties of our optimal threshold group sequential tests and the error spending designs.

6.1.1 Definitions

Recall that we wish to test H_0:$\mu \leq 0$ against the one-sided alternative H_1:$\mu > 0$ with error probabilities α and β at $\mu = 0$ and $\mu = \delta$ respectively. In our notation, μ is

the difference in treatment efficacy between a new treatment and a control and δ is a medically significant difference in efficacy. We can take observations $X_1, X_2, \ldots$, each of which is assumed to be independent and to have a normal distribution with mean μ and known variance σ^2. We shall continue to observe groups of data until either the group sequential boundary is crossed or the total number of observations seen exceeds a fixed threshold value n_T. We do not fix the number of observations taken between each analysis, but instead use a simple probability distribution to model the possible group sizes. We define the random variable N_i to be the sample size at the ith analysis, and we also define the size of the ith group of observations to be M_i. The subscript i is reserved for use as an index which denotes the analysis number under consideration. We still use K for the analysis at which the trial will be forced to terminate, but this is now a random variable rather than a fixed number.

We consider models where the group size at each analysis is an independent and identically distributed random variable drawn from a simple discrete distribution. At each analysis i, the group size M_i can take any one of η possible values, $m(1), \ldots, m(\eta)$. The probability of group i consisting of $m(j)$ observations is defined to be ε_j, that is $\mathbb{P}\{M_i = m(j)\} = \varepsilon_j$ for $j = 1, \ldots, \eta$.

With this definition, we are restricted to models with the same possible group sizes at each analysis. The method we shall present could easily be extended to the more general case where the possible group sizes at different analyses are not the same, or where the probabilities $\{\varepsilon_1, \ldots, \varepsilon_\eta\}$ change from one analysis to the next. We could also generalise this method to allow the group sizes at one analysis to affect the possible group sizes at subsequent analyses.

Following the notation introduced in chapter 5, we write $\mathbf{m}$ for a group size sequence $(m_1, \ldots, m_K)$ and we write $\mathbf{n}$ for the associated sample size sequence. Given the probabilities $\{\varepsilon_1, \ldots, \varepsilon_\eta\}$, the probability of observing a particular group size sequence

$\mathbf{m} = \big(m(j_1), \ldots, m(j_K)\big)$ will be given by

$$\mathbb{P}\Big\{M_1 = m(j_1) \cap \ldots \cap M_K = m(j_K)\Big\} = \varepsilon_{j_1} \ldots \varepsilon_{j_K}.$$

We define $\mathcal{N}$ to be the set of possible cumulative sample sizes which may be observed at some point during the trial. This set will arise as a consequence of the possible group sizes and observation threshold n_T. We shall write its elements as $\mathcal{N} = \big\{n(1), \ldots, n(L)\big\}$, where $n(1) < n(2) < \cdots < n(L)$ and L is the number of possible sample sizes which may arise during the course of the trial. Thus the cumulative sample size at analysis i is a random variable N_i which takes values in $\mathcal{N}$ (although not all the elements of $\mathcal{N}$ are possible values of N_i at any given analysis i).

After seeing $n(l)$ observations with the sum of these values being denoted $S_{n(l)}$, our action will be determined by constants $a(l)$ and $b(l)$, with $a(l) \leq b(l)$, as follows.

If	$S_{n(l)} < a(l)$	STOP, accept H_0,
if	$a(l) < S_{n(l)} < b(l)$	continue to next analysis,
and if	$b(l) < S_{n(l)}$	STOP, reject H_0.

For every l such that $n(l) \geq n_T$, we shall set $a(l) = b(l)$ to ensure that the trial will terminate once the number of observations seen has reached our chosen threshold.

Unlike the random group sequential tests discussed in chapter 5, these boundaries have a single continuation region after a specific number of observations have been seen, regardless of the number of analyses taken to reach that point. This is because once we have seen $n(l)$ observations with sum $S_{n(l)}$ the probability distributions of the next group of observations and of the sum of these observations will be the same regardless of how many analyses have already been carried out. We refer to these tests as threshold group sequential tests. As with the random group sequential tests discussed in chapter 5, applying a threshold group sequential test to any realisation of

the group size sequence will result in a fixed group sequential test. The properties of these fixed group sequential tests, conditional upon an achieved group size sequence, are considered in §6.2.

We find our optimal threshold group sequential tests via the same Bayesian decision theory problem we have used in previous chapters. The formulation has not changed from that discussed in §5.1.2, but we restate the problem in §6.1.2 for convenience. Some changes are necessary to the backward induction algorithm we use to solve the Bayesian problem, and these are discussed in section §6.1.3. We then go on to describe several possible group size models in section §6.1.4.

6.1.2 The Bayes problem

The formulation of our Bayes decision problem has not changed from that defined in §5.1.2, and hence we restate it briefly here. More details are given in the earlier discussion of this problem.

We wish to decide between D_0:'$\mu = 0$' and D_δ:'$\mu = \delta$', where μ is the difference in efficacy between new and control treatments, with positive values of μ implying that the new treatment is superior. We place a prior $\pi(\mu)$ on μ and define the cost of taking one observation to be $c(\mu)$. We also define the loss function $L_2(D, \mu)$ which gives the cost of making a decision to be $L_2(D_0, \delta) = d_\delta$, $L_2(D_\delta, 0) = d_0$ and $L_2(D, \mu) = 0$ otherwise.

Our goal is to find a Bayes rule which corresponds to a threshold group sequential test minimising one of the objective functions defined on page 50. For any of the objective function we wish to minimise, we can construct a decision problem with objective function F_r, prior $\pi(\mu)$, decision costs d_0 and d_δ and constants k_1 and k_2 which has

expected cost given by

$$\begin{aligned}\mathbb{E}\{\text{cost}\} &= \mathbb{E}\{\text{cost of sampling}\} + \mathbb{E}\{\text{cost of decision}\} \\ &= k_1 F_r + k_2 \left(d_\delta \pi(\delta) \mathbb{P}_\delta\{D_0\} + d_0 \pi(0) \mathbb{P}_0\{d_\delta\} \right),\end{aligned}$$

For any given values of d_0 and d_δ, the Bayes rule will have probabilities of making incorrect decisions $\mathbb{P}_\delta\{D_0|d_0, d_\delta\}$ and $\mathbb{P}_0\{D_\delta|d_0, d_\delta\}$, and the Bayes rule will minimise the total expected cost of all decision rules with these probabilities of wrong decisions. Hence, the Bayes rule must minimise F_r amongst all decision rules with these probabilities of wrong decisions.

To identify the optimal threshold group sequential test minimising F_r with error rates α and β, we then search over d_0 and d_δ to find the Bayes rule which has probabilities of wrong decisions $\mathbb{P}_\delta\{D_0|d_0, d_\delta\} = \alpha$ and $\mathbb{P}_0\{D_\delta|d_0, d_\delta\} = \beta$. As with the random group sequential tests found in the previous chapter, we use the iterative method of calculating error probabilities and objective functions described in §5.1.4 as evaluating the error probabilities for each group size sequence $\mathbf{m}$ and weighting the average by the group size model would be inefficient.

6.1.3 Backward induction for the information threshold design

The backward induction algorithms used to solve the Bayes decision theory problems in earlier chapters have all solved problems where the maximum number of analyses, K, is fixed as part of the test design. This is no longer the case, and in this section we shall discuss using the backward induction method to find the optimal decision rule when we replace the maximum number of analyses by the information threshold. Instead of considering our actions at analysis K, then analysis $K-1$ and proceeding backwards to analysis 1, we now work backwards through the set of possible sample sizes $\mathcal{N}$, finding the optimal continuation region for each possible sample size $n(l)$ in decreasing order

$n(L), n(L-1), \ldots, n(1)$. As we noted in the previous section, our action at a specific sample size will be determined only by the number of observations seen and the sum of these observations, not by the number of analyses we have already carried out before reaching that point.

Let $p(\mu|s_{n(l)})$ be the posterior probability for μ after we have seen $n(l)$ observations with sum $s_{n(l)}$. The posterior is given by

$$\begin{aligned} p(\mu|s_{n(l)}) &\propto \pi(\mu) f_\mu(s_{n(l)}) \\ &= \frac{1}{\sqrt{2\pi n(l)\sigma^2}} \exp\left\{ \frac{\left(s_{n(l)} - n(l)\mu\right)^2}{-2n(l)\sigma^2} \right\}, \end{aligned}$$

where $f_\mu(s_{n(l)})$ is the probability density function of $S_{n(l)}$ if the mean of each observation is μ.

Given $n(l)$ observations with the sum of these observations being $s_{n(l)}$ at analysis i, we define the expected cost of stopping and choosing between D_0 and D_δ to be $\gamma(s_{n(l)})$. We also define the expected cost of taking a further group of observations and proceeding optimally once these observations have been seen to be $\beta(s_{n(l)})$.

We first consider the optimal action to take if we have seen $n(L)$ observations. Since $n(L) > n_T$, we shall wish to ensure the termination of the trial after $n(L)$ observations, thus we shall set $a(L) = b(L) = s^*_{n(L)}$, where $s^*_{n(L)}$ is such that

$$\begin{aligned} \mathbb{E}\{\text{cost of } D_0 | S_{n(l)} = s^*_{n(L)}\} &= \mathbb{E}\{\text{cost of } D_\delta | S_{n(L)} = s^*_{n(L)}\} \\ \Rightarrow \qquad d_\delta p(\delta|s^*_{n(L)}) &= d_0 p(0|s^*_{n(L)}). \end{aligned}$$

Solving this equation, we find that

$$s^*_{n(L)} = \frac{\delta n(L)}{2} - \frac{\sigma^2}{\delta} \log\left\{ \frac{d_\delta \pi(\delta)}{d_0 \pi(0)} \right\}. \tag{6.1}$$

We then consider the optimal action to take after seeing $n(L-1)$ observations, then after $n(L-2)$ observations and so on. For each l such that $n(l) \geq n_T$, we set $a(l) = b(l) = s^*_{n(l)}$, where $s^*_{n(l)}$ is found by replacing $n(L)$ with $n(l)$ in equation (6.1).

For each l such that $n(l) < n_T$, we must find $a(l) \leq s^*_{n(l)} \leq b(l)$ to define the continuation region having seen $n(l)$ observations with sum $s_{n(l)}$. Firstly, we find $s^*_{n(l)}$ and evaluate the expected cost of stopping and the expected cost of continuing to the next analysis and proceeding optimally there if we observe $S_{n(l)} = s^*_{n(l)}$. If $\gamma(s^*_{n(l)}) < \beta(s^*_{n(l)})$, we minimise the expected cost of our action by stopping and choosing whichever of D_0 and D_δ has the lower expected cost. In this case, we set $a(l) = b(l) = s^*_{n(l)}$. If, however, $\gamma(s^*_{n(l)}) > \beta(s^*_{n(l)})$, we search for two values of $s_{n(l)}$ such that $\gamma(s_{n(l)}) = \beta(s_{n(l)})$, one value above and one value below $s^*_{n(l)}$. We then set $a(l)$ and $b(l)$ to these values, such that $a(l) < s^*_{n(l)} < b(l)$. In order to carry out this search, we require monotonicity of $\gamma(s_{n(l)}) - \beta(s_{n(l)})$, as discussed on page 29. As with the other Bayes decision problems we have considered, we have not found any examples where this monotonicity does not hold.

Having discussed the searches for the boundary points at each possible sample size we must now discuss the evaluation of $\gamma(s_{n(l)})$ and $\beta(s_{n(l)})$. Before doing so it is necessary to define two pieces of notation. If we have seen $n(l)$ data values and an observed value $s_{n(l)}$ of $S_{n(l)}$ at analysis i we write $n(k)$ for the sample size at analysis $i+1$. Thus k is such that $n(k) = n(l) + m(j)$ for some $j \in \{1, \ldots, \eta\}$. We then define the cumulative distribution function of $S_{n(k)}$ given particular values of $n(k)$ and $s_{n(l)}$ to be

$$F(s_{n(k)}|s_{n(l)}) = \sum_{\mu \in \mathcal{M}} \left\{ p(\mu|s_{n(l)}) g_\mu(s_{n(k)}|s_{n(l)}) \right\},$$

where g_μ is the probability density function of $S_{n(k)}$ given both $s_{n(l)}$ and a specific value of μ, and $\mathcal{M}$ is the set of values of μ upon which the prior $\pi(\mu)$ places positive probability mass.

The expected cost of making a decision at analysis i with $n(l)$ observations yielding a value $s_{n(l)}$ of the summary statistic $S_{n(l)}$ is

$$\gamma(s_{n(l)}) = \min\left\{d_\delta p(\delta|s_{n(l)}), d_0 p(0|s_{n(l)})\right\},$$

which can easily be calculated once the posterior for μ is known. For any l such that $n(l) \geq n_T$, we can not take a further group of observations, but for any l such that $n(l) < n_T$, the expected cost of continuing to the next analysis and acting optimally there is

$$\begin{aligned}\beta(s_{n(l)}) &= \sum_{\mu\in\mathcal{M}}\left[c(\mu)p(\mu|s_{n(l)})\sum_{j=1}^{\eta}\left\{\varepsilon_j m(j)\right\}\right] + \\ \beta &\\ &\sum_{j=1}^{\eta}\left\{\varepsilon_j\int_{\mathbb{R}}\min\left\{\gamma(s_{n(k)}),\beta(s_{n(k)})\right\}dF(s_{n(k)}|s_{n(l)})\right\}, \\ &\sum_{j=1}^{\eta}\left\{\varepsilon_j\int_{\mathbb{R}}\min\left\{\gamma(s_{n(k)}),\beta(s_{n(k)})\right\}dF(s_{n(k)}|s_{n(l)})\right\},\end{aligned}$$

where $n(k) = n(l) + n(j)$.

In order to evaluate the integral over $s_{n(k)}$ numerically we need to know the expected cost of taking the optimal action for a grid of values of $S_{n(k)}$ for each possible $n(k)$. In order to do this, we find the continuation regions sequentially, starting with the continuation region for the largest $n(l)$ which is less than n_T and proceeding backwards to $n(1)$. At each stage, once $a(l)$ and $b(l)$ have been found, we evaluate $\beta(s_{n(l)})$ on a grid of values of $s_{n(l)}$ and use these values in subsequent numerical integrations.

This process is very similar to the backward induction algorithms which we used in earlier chapters. The key difference is that we no longer consider the action at the final analysis and work backwards to earlier analyses. Instead we first consider our action for the largest possible cumulative sample size that may occur, $n(L)$, and proceed to find the optimal action for smaller possible sample sizes in decreasing order. This is summarised in algorithm 4.

Algorithm 4: Asymmetric maximum information algorithm

- For $l = L, \ldots, 1$:
  - Find $s^*_{n(l)}$ such that $d_\delta p(\delta|s^*_{n(l)}) = d_0 p(0|s^*_{n(l)})$.
  - If $n(l) \geq n_T$ or if $\gamma(s^*_{n(l)}) \leq \beta(s^*_{n(l)})$,
    - set $a(l) = b(l) = s^*_{n(l)}$.
  - Otherwise,
    - find $a(l)$ such that $\gamma(a(l)) = \beta(a(l))$, with $a(l) < s^*_{n(l)}$,
    - find $b(l)$ such that $\gamma(b(l)) = \beta(b(l))$, with $b(l) > s^*_{n(l)}$,
    - evaluate $\beta(s_{n(l)})$ on a grid of values of $s_{n(l)}$ from $a(l)$ to $b(l)$.

Comparing this algorithm to that for the random group sequential tests of chapter 5 on page 83, it can be seen that considering what action to take after a given number of observations, regardless of the number of analyses seen, simplifies the backward induction process.

6.1.4 Some group size models

In this section we define several group size models which we shall use in the rest of this chapter to examine the properties of the threshold group sequential tests. Recall that we define our group size models by a set of possible group sizes $\{m(1), \ldots, m(\eta)\}$ with associated probabilities $\varepsilon_j = \mathbb{P}\{M_i = m(j)\}$ and the information threshold n_T. In all of the models we consider the sets of possible group sizes and associated probabilities do not depend on the number of analyses already carried out, thus the sets $\{m(1), \ldots, m(\eta)\}$ and $\{\varepsilon_1, \ldots, \varepsilon_\eta\}$ are the same for each group size. We will force termination of the trial at the first analysis where n_T or more observations have been seen, if the trial has not already terminated by this point.

For each model, specifying $\{m(j), \varepsilon_j; j = 1, \ldots, \eta\}$, and n_T defines the model entirely. We define $\mathcal{K}$ to be the set of possible values of K, the analysis at which the trial will be forced to terminate and $\mathcal{N}$ to be the set of possible sample sizes which may be observed in the course of the trial. These sets are noted below in the description of each model in order to make comparisons of the models easier.

For all the examples considered in this chapter, we shall be testing H_0:$\mu \leq 0$ against H_1:$\mu > 0$ with type I error probability $\alpha = 0.05$ at $\mu = 0$ and equal type II error probability $\beta = 0.05$ at $\mu = \delta = 0.25$. The variance of each observation is $\sigma^2 = 1.0$, leading to the number of observations required for a non-sequential test being $n_{fix} = 43.30$ to two decimal places.

Model 1

m	10	15	20
$\mathbb{P}\{M = m\}$	0.25	0.50	0.25

$n_T = 45$ $\quad \mathcal{K} = \{3, 4, 5\}$ $\quad \mathcal{N} = \{10, 15, \ldots, 60\}$

Model 1 is a simple model which represents an expectation that the group sizes will mostly follow an anticipated schedule, but with some possibility of groups being substantially larger or smaller than planned. This model is similar in spirit to the sample size models 3 and 6 in chapter 5.

Model 2

m	5	10	15	20	25
$\mathbb{P}\{M = m\}$	0.1	0.2	0.4	0.2	0.1

$n_T = 45$ $\quad \mathcal{K} = \{2, 3, \ldots, 9\}$ $\quad \mathcal{N} = \{5, 10, \ldots, 65\}$

Model 3

m	10	15	20
$\mathbb{P}\{M = m\}$	1/3	1/3	1/3

$n_T = 45$ $\quad \mathcal{K} = \{3, 2, 5\}$ $\quad \mathcal{N} = \{10, 15, \ldots 60\}$

Models 2 and 3 can both be thought of as extensions to model 1 with greater uncertainty as to the group sizes. In model 2 this uncertainty is expressed as a greater range of possible group sizes, while in model 3 we have the same possible group sizes as model 1 but have equal probabilities of each group size occurring.

Model 4

m	10	11	$\cdots$	20
$\mathbb{P}\{M = m\}$	1/11	1/11	$\cdots$	1/11

$n_T = 45$ $\quad \mathcal{K} = \{3, 4, 5\}$ $\quad \mathcal{N} = \{10, 11, \ldots, 64\}$

Model 4 can also be considered as an extension to model 1 where there is less certainty as to the possible group sizes. The range of possible group sizes is the same as in model 1, but we now consider groups which can be any size between 10 and 20 observations. If more frequent analyses with smaller group sizes were considered, this style of model could approach the state of discretised fully sequential monitoring, where the data could be analysed after every observation. We have assigned equal probability to each possible group size in this model, but an obvious alternative would be to use a Poisson distribution of some kind, truncating the possible group sizes and renormalising the probabilities.

Model 5

m	5	10	15	20	25
$\mathbb{P}\{M = m\}$	0.1	0.2	0.4	0.2	0.1

$n_T = 40$ $\quad \mathcal{K} = \{2, 3, \ldots, 8\}$ $\quad \mathcal{N} = \{5, 10, \ldots, 60\}$

Model 5 is similar to model 2, but with a lower value of the observation threshold n_T. Models 1 to 4 all have $n_T > n_{fix}$, ensuring that the number of observations taken would be sufficient to carry out the equivalent fixed-sample test. Model 5, however, has $n_T < n_{fix}$ and thus includes the possibility of being forced to terminate the trial when the number of observations taken would be insufficient to carry out a fixed sample test with the desired error rates. This enables us to consider the effects of under-powered sample size sequences on out optimal tests and the error spending method. In trials where the design calls accumulation of data until a fixed level of information has been reached, this can occur due to non-clinical reasons, such as lack of funding.

6.2 Conditional properties of the information threshold tests

The main motivation for our optimal threshold group sequential tests is to assess the performance of the error spending tests with respect to their expected sample sizes, but first we wish to consider the properties of the optimal tests in their own right. As with the random group sequential tests of chapter 5, our threshold group sequential tests can be considered as a set of fixed group sequential tests, of the type found in chapters 3 and 4. Each element of the set corresponds to a fixed group sequential boundary for a specific sequence of group sizes $\mathbf{m} = \big(m(j_1), \ldots, m(j_K)\big)$, where we note that while we still use K to denote the maximum number of analyses, it is no longer fixed as

has been the case in previous chapters. The number of observations seen by analysis i determines the upper and lower bounds of the continuation region which we have found by the backwards induction algorithm discussed in §6.1.3.

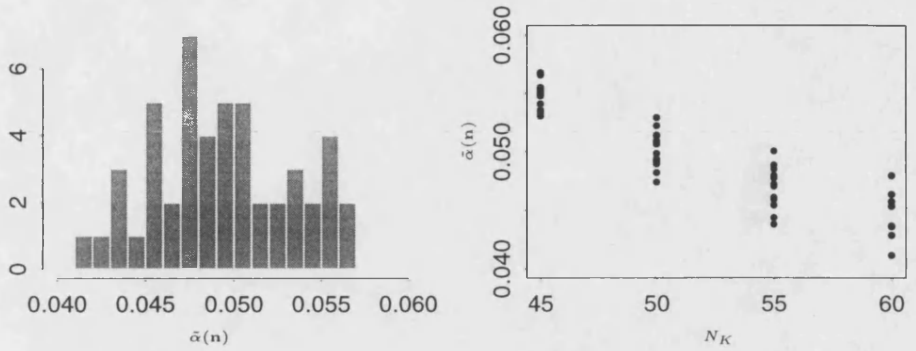
In this section, we consider the properties of the fixed group sequential tests which comprise the optimal threshold group sequential test. Any specific group size sequence $\mathbf{m}$ uniquely determines a sample size sequence $\mathbf{n}$ and we shall discuss the properties of the tests conditional upon the sample size sequence following the discussions in chapter 5. In §6.2.1, we discuss the achieved conditional error probabilities, in §6.2.2 we discuss the conditional and average objective function values achieved by the threshold group sequential test, and in §6.2.3 we discuss the risks of the decision rules corresponding to our optimal tests.

6.2.1 Achieved error probabilities

Our optimal threshold group sequential tests can be viewed as a collection of fixed group sequential tests, one for each sample size sequence which may occur during the course of the clinical trial. We define $\widetilde{\alpha}(\mathbf{n})$ to be the achieved type I error probability of the fixed group sequential test obtained by applying the threshold group sequential test to sample size sequence $\mathbf{n}$, with an analogous definition for the achieved type II error probability $\widetilde{\beta}(\mathbf{n})$. Due to the symmetric nature of the examples discussed in this chapter, with overall error probabilities $\alpha = \beta = 0.05$, for any given group size model and sample size sequence $\mathbf{n}$ an optimal threshold group sequential test will have $\widetilde{\alpha}(\mathbf{n}) = \widetilde{\beta}(\mathbf{n})$.

Figure 6-1 shows the values of $\widetilde{\alpha}(\mathbf{n})$ achieved by the threshold group sequential test optimised for objective function F_1 and following models 1 and 5. These results show a range of values of $\widetilde{\alpha}(\mathbf{n})$, from 0.041 to 0.057 for model 1 and from 0.033 to 0.061 for model 5. The most striking difference between the results for model 1 and those for

MODEL 1



MODEL 5

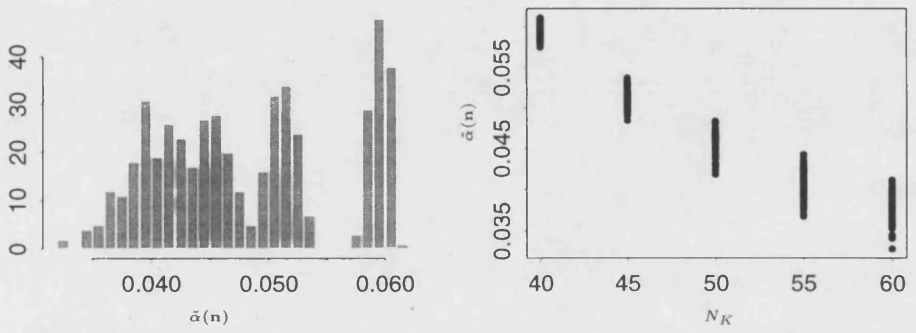


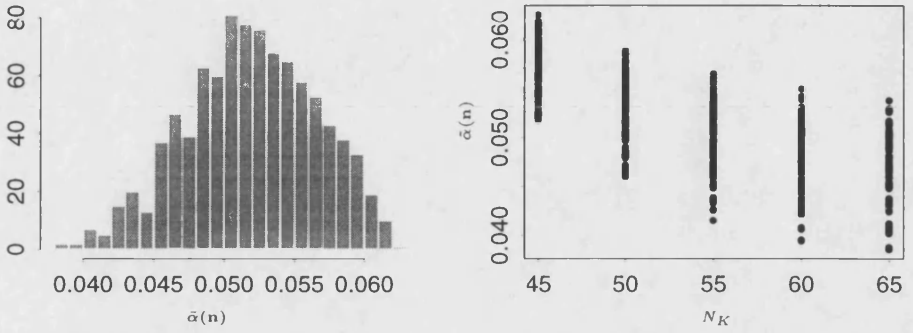
Figure 6-1: *Achieved type I error probability conditional upon sample size sequences for two optimal threshold group sequential tests. The group sizes follow models 1 and 5 and the tests are optimised for* F_1. *Overall type I and type II error probabilities are 0.05.*

the random group sequential tests is that the values of $\tilde{\alpha}(\mathbf{n})$ no longer fall into discrete clusters, depending upon the final sample size n_K. The scatter plot shows that there is still a clear link between the maximum sample size and the achieved error probability, however the link is no longer as strong as in the random group sequential tests we studied in chapter 5.

There are two factors which make the effect of the maximum sample size on achieved error probability less striking for threshold group sequential tests than for the random group sequential tests we considered in chapter 5. Firstly, for any specific value n_K of N_K there are more possibilities for the sample size sequence with that n_K. In most cases, there will be sequences with different numbers of analyses leading to the same n_K. This increases the range of different values of $\tilde{\alpha}(\mathbf{n})$ to be seen for any given maximum sample size. Secondly, it was noted in chapter 5 (page 95) that the clusters of $\tilde{\alpha}(\mathbf{n})$ values became less distinct as the maximum sample size increased. For model 1, all the maximum sample sizes are greater than the fixed sample size n_{fix}, while for the random group sequential tests there were sample size sequences with $n_K < n_{fix}$. Achieved values of $\tilde{\alpha}(\mathbf{n})$ for the threshold group sequential test following model 5 and optimised for F_1 are shown in the lower portion of figure 6-1. Model 5 has $n_T < n_{fix}$, and the clustering of $\tilde{\alpha}(\mathbf{n})$ values observed in chapter 5 is more noticeable here than in the results for the threshold group sequential test following model 1.

Figure 6-2 shows the range of $\tilde{\alpha}(\mathbf{n})$ value achieved by the threshold tests optimised for F_1 following group size models 2 and 3. In a sense, both these models are generalisations of model 1 with a greater degree of uncertainty as to the possible group sizes. Model 2 has a greater range of possible group size, but still concentrates probability mass on each group having 15 observations. In contrast, model 3 has the same three possible group sizes as model 1, but places equal probability mass upon each. Model 2 has many more possible sample size sequences than model 1, and this results in the histogram revealing a much smoother distribution for $\tilde{\alpha}(\mathbf{n})$. The scatter plot for model 2 shows a

MODEL 2



MODEL 3

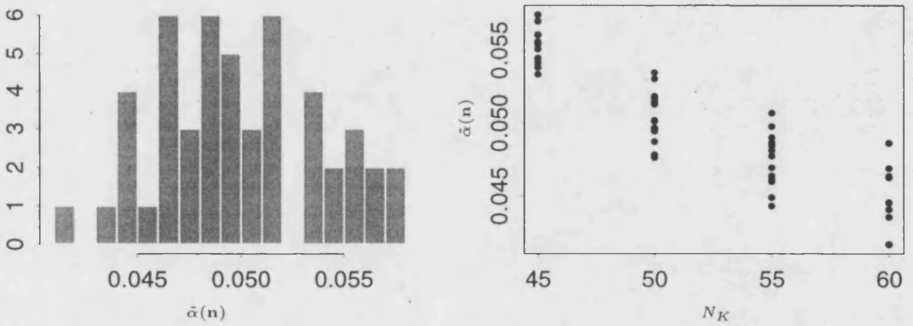


Figure 6-2: *Achieved type I error probability conditional upon sample size sequences for two optimal threshold group sequential tests. The group sizes follow models 2 and 3 and the tests are optimised for* F_1. *Overall type I and type II error probabilities are 0.05.*

Model	$\widetilde{\alpha}(\mathbf{n})$ ranges			
	F_1	F_{21}	F_{31}	F_4
1	0.041 – 0.057	0.039 – 0.059	0.035 – 0.063	0.039 – 0.059
2	0.039 – 0.063	0.036 – 0.066	0.029 – 0.077	0.036 – 0.066
3	0.042 – 0.057	0.039 – 0.059	0.035 – 0.063	0.039 – 0.059
4	0.043 – 0.058	0.040 – 0.061	0.035 – 0.066	0.040 – 0.061
5	0.033 – 0.061	0.032 – 0.062	0.030 – 0.065	0.033 – 0.062

Table 6.1: *Ranges of achieved conditional error probabilities for optimal threshold group sequential tests. Values are for tests following group size models 1 to 6, and minimising objective functions F_1, F_{21}, F_{31} and F_4. All tests had average type I and type II error probabilities $\alpha = \beta = 0.05$.*

link between the maximum sample size and achieved $\widetilde{\alpha}(\mathbf{n})$, but the link is now relatively weak compared to the variation in $\widetilde{\alpha}(\mathbf{n})$ for any fixed n_K. The overall range of $\widetilde{\alpha}(\mathbf{n})$ values is also slightly larger for model 4 than for model 1, while the range of conditional error rates for model 3 is almost exactly the same as for model 1.

Similar patterns were seen in threshold group sequential tests optimised for F_{21}, F_{31}, and F_4 and for other group size models. Ranges of $\widetilde{\alpha}(\mathbf{n})$ values are shown in table 6.1 for tests following all the group size models we defined in §6.1.4 and optimised for all of the objective functions under consideration. For each of these models, the narrowest range of $\widetilde{\alpha}(\mathbf{n})$ values is for the tests minimising F_1, while the tests minimising F_{31} have the widest ranges of conditional error probabilities. From the differences between these results for models 1 and 2 we can see that increasing the range of possible group sizes increases the range of values of $\widetilde{\alpha}(\mathbf{n})$ observed. However, comparing the results for models 1 and 3 show that altering the probabilities of the different group sizes occurring has little effect upon the ranges of conditional error probabilities. From the results for model 4, we can see that allowing each possible group size between 10 and 20 observations has altered the range of conditional error rates very little when compared with model 1, which has the same range of group sizes but only allows 10, 15, or 20 observations per group. Finally, by comparing the results for model 2 with those for

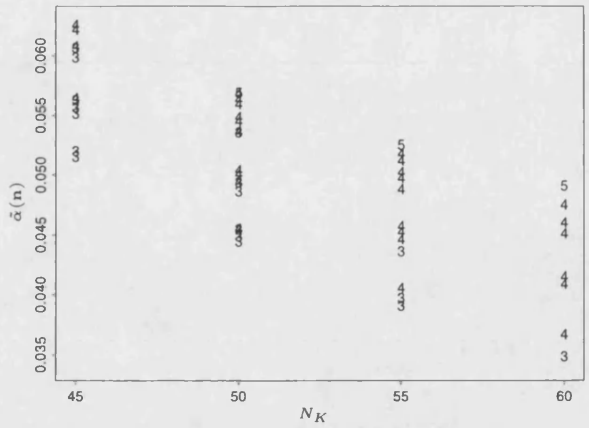


Figure 6-3: *Achieved type I error probability conditional upon sample size sequences for an optimal threshold group sequential test. The group sizes follow model 1 and the test is optimised for F_{31}. Overall type I and type II error probabilities are 0.05. Each group size sequence had the number of analyses indicated by the plotting character.*

model 5 we can see that lowering the information threshold has had a small effect on the conditional error rates, and that the changes have mostly been to make the error rates more conservative.

Within the range of $\tilde{\alpha}(\mathbf{n})$ values for any specific value of N_K, the number of analyses carried out strongly affects the achieved $\tilde{\alpha}(\mathbf{n})$. Figure 6-3 shows the achieved $\tilde{\alpha}(\mathbf{n})$ values for the test following model 1 optimised for objective function F_{31}, plotted against the achieved maximum sample size. The plotting characters are the value of K for each sample size sequence $\mathbf{n}$. For example, if the trial is forced to terminate after a total of 55 observations have been taken, there could have been three, four, or five analyses. The figure shows that achieved value of $\tilde{\alpha}(\mathbf{n})$ will be larger if there have been four analyses than if three analyses have been carried out, and larger still if there have been five analyses for any particular value of N_K. This is similar to the well-known property of error inflation caused by increasing the number of analyses which has been discussed by several authors, including Armitage, McPherson & Rowe (1969).

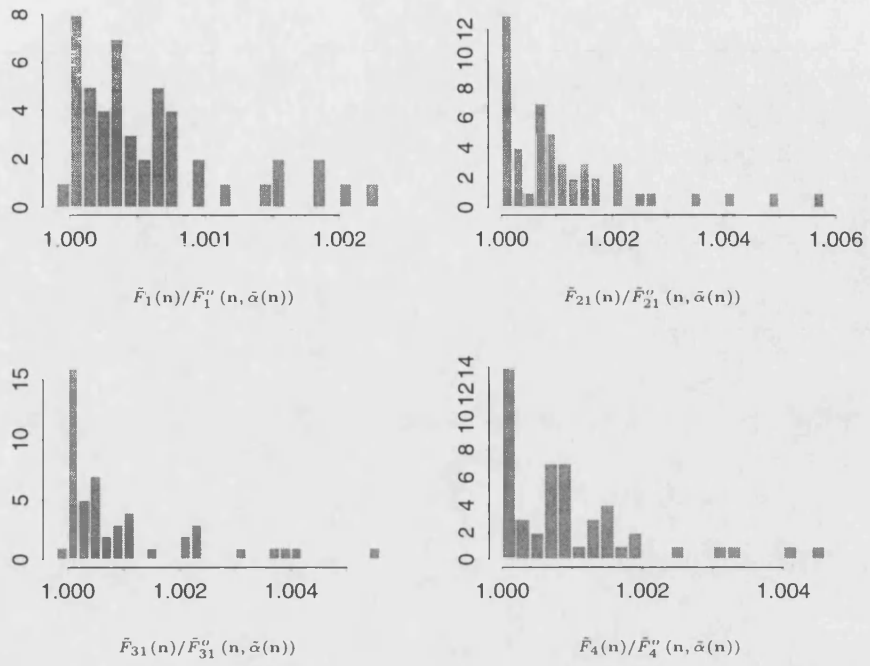


Figure 6-4: *Plotted values are ratios of objective function values conditional upon the sample size sequence* $\mathbf{n}$ *to the values optimal for that* $\mathbf{n}$ *and* $\tilde{\alpha}(\mathbf{n})$. *All tests follow group size model 1 and objective functions* F_1, F_{21}, F_{31}, *and* F_4 *are minimised. Average error rates are* $\alpha = \beta = 0.05$.

6.2.2 Achieved objective function values

We define $\widetilde{F}_r(\mathbf{n})$ to be the value of objective function F_r achieved by the threshold group sequential test when the observed sample size sequence is $\mathbf{n}$. We also define $\widetilde{F}_r^o(\mathbf{n}, \tilde{\alpha}(\mathbf{n}))$ to be the optimal value of objective function F_r among all fixed group sequential tests which have sample size sequence $\mathbf{n}$ and type I and type II error probabilities equal to $\tilde{\alpha}(\mathbf{n})$, the error probability of the optimal threshold group sequential test conditional upon $\mathbf{n}$. The averages of these values are denoted by $\widetilde{F}_r$ and $\widetilde{F}_r^o$ respectively. Figure 6-4 shows histograms of the ratios $\widetilde{F}_r(\mathbf{n})/\widetilde{F}_r^o(\mathbf{n}, \tilde{\alpha}(\mathbf{n}))$ for optimal threshold group sequential tests following group size model 1 and minimising F_1, F_{21}, F_{31}, and F_4. These results show that the achieved conditional objective function values are close

Model	r	$\widetilde{F}_r$	$\widetilde{F}_r^o$	$\widetilde{F}_r/n_{fix}$
1	1	35.81	35.80	0.827
	21	28.20	28.19	0.651
	31	19.13	19.12	0.442
	4	28.10	28.08	0.649
2	1	36.04	36.02	0.833
	21	28.55	28.51	0.659
	31	19.66	19.64	0.454
	4	28.49	28.46	0.657
3	1	35.81	35.80	0.827
	21	28.21	28.18	0.652
	31	19.19	19.17	0.443
	4	28.13	28.11	0.650

Model	r	$\widetilde{F}_r$	$\widetilde{F}_r^o$	$\widetilde{F}_r/n_{fix}$
4	1	35.50	35.49	0.820
	21	27.71	27.70	0.640
	31	18.74	18.73	0.433
	4	27.69	27.68	0.640
5	1	37.88	37.86	0.875
	21	31.00	30.96	0.716
	31	21.59	21.56	0.500
	4	30.57	30.54	0.706

Table 6.2: *Minimum values of objective functions F_1, F_{21}, F_{31}, and F_4 for threshold group sequential tests following models 1 to 5. Values are averaged over the sample size model chosen, and all tests have error probabilities $\alpha = \beta = 0.05$.*

to the minimum possible for matching group sizes and error rates in each case. For model 1, all the ratios $\widetilde{F}_r(\mathbf{n})$ / $\widetilde{F}_r^o(\mathbf{n}, \widetilde{\alpha}(\mathbf{n}))$ were less than 1.006, indicating that the objective function values for all the possible sample size sequences were close to being conditionally optimal. The ratios for model 3 were similar, while those for model 2 were larger, up to approximately 1.03. Maximum values of the ratio $\widetilde{F}_r(\mathbf{n})$ / $\widetilde{F}_r^o(\mathbf{n}, \widetilde{\alpha}(\mathbf{n}))$ for model 5 were very close to those for model 2.

Overall objective function values averaged by the group size model are given in table 6.2 for both the optimal threshold group sequential tests and the tests optimised conditionally on $\mathbf{n}$ and $\widetilde{\alpha}(\mathbf{n})$. In all cases, the value of $\widetilde{F}_r$ is only slightly larger than that of $\widetilde{F}_r^o$, indicating that knowledge of the sample size sequence $\mathbf{n}$ does not significantly alter the optimal objective function vale which can be achieved.

Comparing the results for the different models, we can see that the results are fairly robust to alterations in the group size models. Keeping the same possible group sizes but altering the probability of each occurring makes only a very small difference as

shown by the differences between the results for models 1 and 3. By comparing the results for models 1 and 2, we can see that increasing the range of possible group sizes has a larger effect, as does lowering the observation threshold n_T, but these differences are still small. By keeping the range of group sizes the same but allowing more values within that range, as we do when going from model 1 to model 4, we see a small improvement in the optimum objective function values.

6.2.3 Bayes risk of the threshold group sequential tests

We have considered the performance of our optimal threshold group sequential tests in §6.2.1 and §6.2.2 with respect to the frequentist properties of error probability and expected sample size. We now consider the efficiency of the corresponding Bayesian decision rules by examining the expected cost of these rules.

In §6.1.2 we defined our decision problem, which is to choose between D_0: '$\mu = 0$' and D_δ: '$\mu = \delta$', with a prior $\pi(\mu)$ on μ. The costs of incorrectly deciding D_δ and D_0 are d_0 and d_δ respectively. We write $\mathbf{d} = (d_0, d_\delta)$ for the decision cost vector, and let $\mathbf{d}_{opt}$ be the value of $\mathbf{d}$ which gives the Bayes rule corresponding to our optimal threshold group sequential test. That is, $\mathbf{d}_{opt}$ is the decision cost vector which gives a Bayes rule with the probabilities of incorrectly deciding D_δ and D_0 to be α and β respectively. We then write $\mathcal{R}_r^{(T)}(\mathbf{d}_{opt})$ for the Bayes risk of the decision rule which corresponds to our optimal random group sequential test optimising objective function F_r. The risk of a decision rule is simply the expected cost of the rule, and our problem has been constructed to have

$$\mathcal{R}_r^{(T)}(\mathbf{d}_{opt}) = k_1 \widetilde{F}_r + k_2 \left(d_\delta \pi(\delta) \mathbb{P}_\delta\{D_0\} + d_0 \pi(0) \mathbb{P}_0\{d_\delta\} \right), \tag{6.2}$$

as discussed in §6.1.2, where F_r is the objective function that the Bayes rule has been designed to minimise; the values of the constants k_1 and k_2 depend upon which objective

function we wish to minimise and are listed on page 78. We also define the risk of this rule, conditional upon an observed sample size sequence $\mathbf{n}$ to be

$$\mathcal{R}_r^{(T)}(\mathbf{n}, \mathbf{d}_{opt}) = k_1 \widetilde{F}_r(\mathbf{n}) + k_2 \left(d_\delta \pi(\delta) \mathbb{P}_\delta\{D_0|\mathbf{n}\} + d_0 \pi(0) \mathbb{P}_0\{d_\delta|\mathbf{n}\} \right). \qquad (6.3)$$

If we knew in advance what sample size sequence was to occur in the trial, the same decision cost vector would give a different Bayes rule to that obtained as a special case of the random groups Bayes rule. We can easily calculate this rule using the backward induction algorithm from §6.1.3, optimising the objective function of our choice. Equation (6.3) can then be used to evaluate the risk of this Bayes rule, which we denote $\mathcal{R}_r^{(F)}(\mathbf{n}, \mathbf{d}_{opt})$.

Figure 6-5 shows the values of $\mathcal{R}_r^{(T)}(\mathbf{n}, \mathbf{d}_{opt})/\mathcal{R}_r^{(F)}(\mathbf{n}, \mathbf{d}_{opt})$ for the Bayes rule following group size model 1 and optimising each of F_1, F_{21}, F_{31}, and F_4. Unlike the values in figure 6-4, which were ratios of expected sample sizes for tests which had been forced to have matching error rates, the results in figure 6-5 are risk ratios between decision rules which do not have matching error rates but have the same costs for wrong decisions. The risks take both expected sample sizes and the error probabilities into account. We are comparing the expected costs of the overall Bayes procedure conditional upon a particular sample size sequence $\mathbf{n}$ to the optimal Bayes rule for that $\mathbf{n}$ given the same decision cost vector $\mathbf{d}_{opt}$. From figure 6-5, it is clear that the conditional risks $\mathcal{R}_r^{(T)}(\mathbf{n}, \mathbf{d}_{opt})$ are very close to the lowest possible risks for the same decision cost vector $\mathbf{d}_{opt}$ and sample size sequence $\mathbf{n}$. This indicates that the Bayes rule defined by the decision cost vector $\mathbf{d}_{opt}$ is very close to optimal for each possible sample size sequence individually, as well as having the minimum risk when averaged over the sample size model. Similar patterns were observed for the other sample size models studied.

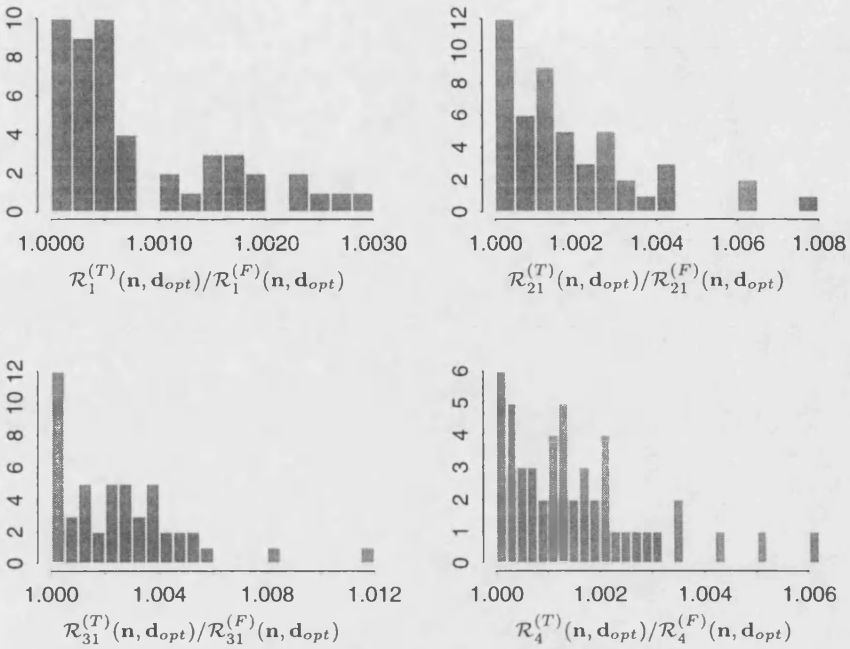


Figure 6-5: *Efficiency of the Bayes rules minimising objective functions* F_1, F_{21}, F_{31}, *and* F_4. *Displayed values are* $\mathcal{R}_r^{(T)}(\mathbf{n}, \mathbf{d}_{opt})$ / $\mathcal{R}_r^{(F)}(\mathbf{n}, \mathbf{d}_{opt})$. *All risks are calculated using the decision cost vector* $\mathbf{d}_{opt}$, *which gives a Bayes rule corresponding to the optimal threshold group sequential test with error rates* $\alpha = \beta = 0.05$ *and following group size model 1.*

6.2.4 Conclusions

We have considered the behaviour of the optimal threshold group sequential tests when applied to a range of group size models. Many other models would be possible and in particular it must be noted that some of the group size models we have considered have possible maximum sample sizes considerably larger than the fixed sample size n_{fix}, however it would be straightforward to construct models which avoided this. For example, we could define models which had different possible group sizes if we were near the fixed sample size; in practice this would correspond to a trial protocol specifying that under these conditions the next analysis would be carried out sooner than had

been the case during the earlier stages of the trial.

Our main motivation in developing the optimal threshold group sequential tests was to provide a tool to assess the performance of the error spending tests, which we shall do in the following section. However, the optimal threshold group sequential tests are useful designs in their own right, provided that a sensible group size model can be formulated. Comparing the results for models 1 and 2 has shown that increasing the range of possible group sizes at each analysis increases the range of values of $\widetilde{\alpha}(\mathbf{n})$ which may be observed and slightly decreases the achieved efficiency. Thus, by controlling the range of possible group sizes we can restrict the variation in conditional error probabilities to a tolerable range. Comparisons between models 1 and 3 have shown that altering the probabilities of the possible group sizes has little effect upon the properties of the test. Models 1 and 4 have the same range of possible group sizes but model 4 permits many more values of M_i in that range. Comparing the results for these models has shown a very light increase in the range of $\widetilde{\alpha}(\mathbf{n})$ values and a slight increase in the efficiency of the tests. Hence, it seems it would be possible to construct models which permitted near-continuous monitoring of the data and to design very flexible trials by using these models.

From the results discussed in §6.2.1 — §6.2.3 and those from other models we have used, it seems that the achieved values of $\widetilde{\alpha}(\mathbf{n})$ and $\widetilde{F}_r(\mathbf{n})$ are somewhat less variable for the threshold group sequential than for the random group sequential designs of chapter 5. This can partly be attributed to the fact that the group size models used in the threshold designs have less variable maximum sample sizes than the sample size models introduced in §5.1.5. Another reason is that few of the maximum sample sizes in the optimal threshold group sequential tests are less than n_{fix}, while in the random group sequential tests we studied models with many values of N_K which were below the sample size required for the equivalent non-sequential design.

6.3 Performance of the error spending method

We now consider the performance of the error spending tests defined by the γ-family and ρ-family error spending functions when applied to the group size models introduced in §6.1.4. The properties of the error spending tests are compared to those of the threshold group sequential tests, firstly with regard to deviations from the nominal error probabilities in §6.3.1, then with respect to the efficiency of the tests in §6.3.2. We do not consider the Δ-family tests as a fixed maximum number of analyses is integral to the definition of the Δ-family tests, and thus applying these tests to the group size models we consider in this chapter would be impractical.

6.3.1 Deviations from the nominal error rates

The error spending tests preserve type I error at exactly the nominal level, but the type II error varies with deviations from the planned group size sequence. Error spending tests were designed for testing $H_0{:}\mu \leq 0$ against $H_1{:}\mu > 0$ with equal type I and type II error probabilities 0.05 at $\mu = 0$ and $\mu = 0.25$. For each of the group size models we defined in 6.1.4 the most likely schedule of analyses is for there to be three groups of 15 observations each, each observation being independent and identically $N(\mu, 1.0)$ distributed. Parameter values $\gamma = -3.24$ and $\rho = 2.71$ result in tests meeting these design criteria. The achieved type II error probabilities of the γ-family tests conditional upon the achieved sample size sequence are shown in figure 6-6 for group size models 1, 2, and 5. For each example we have seen, the type II errors achieved by the γ- and ρ-family error spending tests has been very similar for each sample size sequence $\mathbf{n}$, although those for the ρ-family tests have been slightly larger on average.

It is immediately noticeable that almost the majority of the deviation from the nominal error is downward; that is $\widetilde{\beta}(\mathbf{n}) < \beta$ for most sample size sequences. This is because the error spending test have been designed for a sample size sequence with maximum

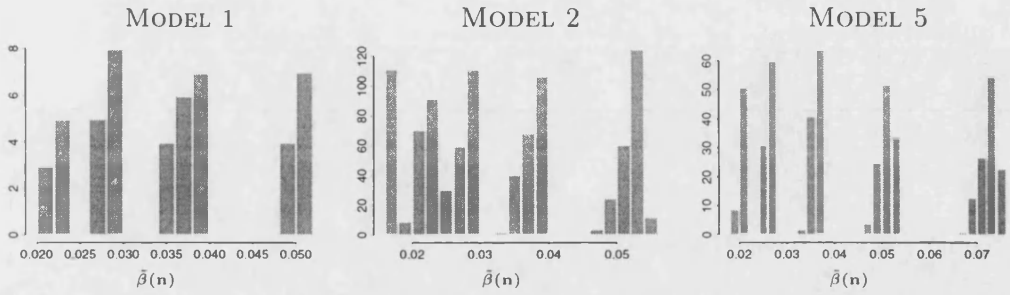


Figure 6-6: *Achieved type II error probability conditional upon sample size sequences for γ-family tests following group size models 1, 2, and 5. All tests were designed for three groups of 15 observations, with equal type I and type II error probabilities $\alpha = \beta = 0.05$.*

sample size equal to the value of n_T in the group size models, which exceeds the sample size for the equivalent non-sequential design in most cases. Model 5 is the exception to this generalisation, as in this model $n_T = 40$, while $n_{fix} = 43.30$. The error spending tests for model 5 have still been designed for a total of 45 observations in three equal groups, and thus in this case there are some sample size sequences where the error spending test is forced to terminate before the planned sample size has been reached. With as few as 5 observations less than planned, the error spending tests see type II error probability as large as 0.075, 50% greater than the intended error probability. For models 1 and 2 there is a small amount of upwards deviation in type II error conditional upon the observed sample size sequence. This error inflation occurs when we reach the target information threshold exactly, but take more than the planned number of analyses during the trial. For example, in model 2 we can take as many as nine analyses to reach our planned maximum sample size of 45 observations. The achieved values of $\widetilde{\beta}(\mathbf{n})$ are mostly clearly divided into clusters, depending upon the achieved value of n_K, although when the group sizes follow model 2 there is some overlapping of the clusters for the larger values of N_K.

To examine the overall variability in error probabilities, incorporating both type I and type II error, we define the total deviation from the nominal error rates conditional

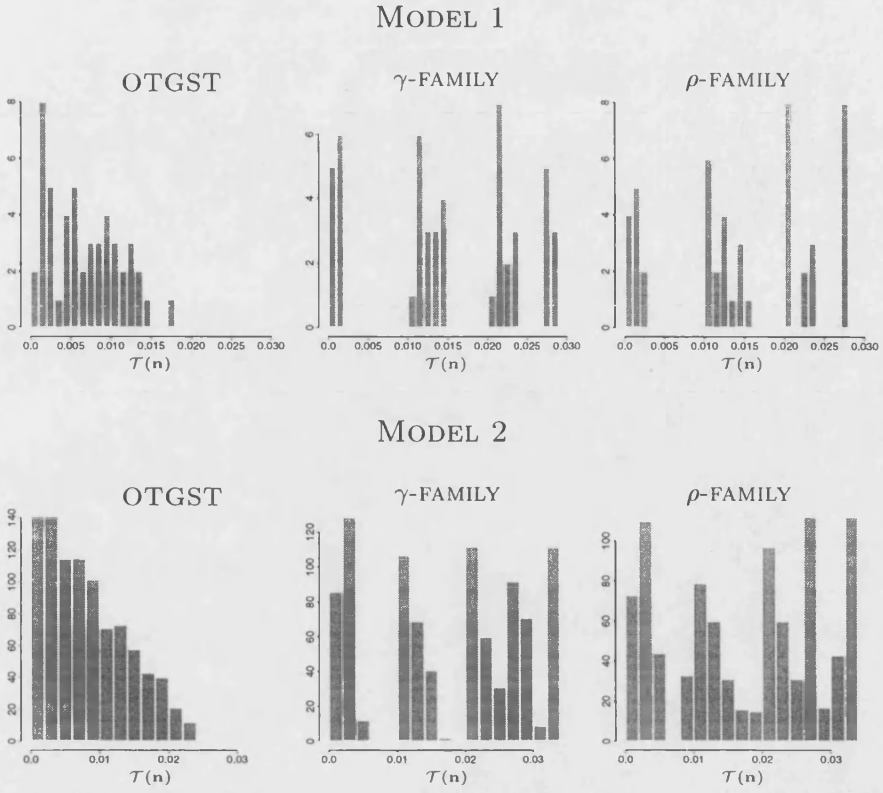


Figure 6-7: *Deviation from the nominal error rates conditional upon sample size sequences for error spending tests following models 1 and 2. Values for the optimal threshold group sequential tests (marked* OTGST*) are shown for comparison. The error spending tests were designed for three groups of 15 observations, with equal type I and type II error probabilities* $\alpha = \beta = 0.05$*. The optimal tests have average error probabilities* $\alpha = \beta = 0.05$.

upon sample size sequence $\mathbf{n}$ to be $\mathcal{T}(\mathbf{n}) = |\widetilde{\alpha}(\mathbf{n}) - \alpha| + |\widetilde{\beta}(\mathbf{n}) - \beta|$. Figure 6-7 shows the deviations from the nominal error rates for the optimal threshold group sequential test and both families of error spending tests when applied to group size models 1 and 2. The threshold group sequential tests have been optimised for objective function F_1. It is clear that the deviations for the error spending tests are, for some sample size sequences, much larger than the deviations for the threshold group sequential test. This has been true for other examples not shown here when the optimal tests have been designed to minimise objective functions F_1, F_{21}, or F_4. Tests designed to minimise F_{31} have shown a greater range of values of $\mathcal{T}(\mathbf{n})$; in some cases the range of deviation from

Model		γ	ρ
1	1	102.6	102.7
	21	103.9	105.0
	31	107.9	111.5
	4	103.6	104.8
2	1	103.3	103.4
	21	104.7	105.8
	31	108.5	111.9
	4	104.3	105.4
3	1	103.0	103.1
	21	104.4	105.5
	31	108.5	112.2
	4	104.0	105.2

Model		γ	ρ
4	1	103.6	103.7
	21	105.4	106.6
	31	110.1	114.3
	4	104.9	106.2
5	1	100.9	101.0
	21	101.2	101.6
	31	102.6	104.2
	4	101.1	101.7

Table 6.3: *Tabulated values are average risks of threshold group sequential tests based on the γ-family and ρ-family designs, given as percentages of the risk of the relevant optimal threshold group sequential test. All risks are evaluated using the cost of a wrong decision from the appropriate optimal threshold group sequential test.*

the nominal error rates α and β for the tests optimising F_{31} was almost as large as that of the error spending tests.

6.3.2 Efficiency of the error spending method

Since the optimal threshold group sequential tests and error spending tests have different conditional error probabilities for any given sample size sequence $\mathbf{n}$, we do not use expected sample size as an efficiency criterion. Instead we use the risk or expected cost of the threshold group sequential tests and the error spending tests to compare the efficiencies of these designs. Recall from §6.1.2 that we define D_0 and D_δ are the decisions '$\mu = 0$' and '$\mu = \delta$', and d_δ and d_0 are the costs of incorrectly making these decisions. We refer to $\mathbf{d} = (d_o, d_\delta)$ as a decision cost vector and for any given particular combination of group size model and objective function of interest there is a value $\mathbf{d}_{opt}$ of $\mathbf{d}$ which gives the Bayes rule corresponding to our optimal threshold

group sequential test. Using an equivalent equation to (6.3), we define the risk of the γ-family error spending test conditional upon sample size sequence $\mathbf{n}$ to be

$$\mathcal{R}_r^{(\gamma)}(\mathbf{n}, \mathbf{d}_{opt}) = k_1 F_r + k_2 \left(d_0 \mathbb{P}_0\{D_{\delta|\mathbf{n}}\} + d E_\delta \mathbb{P}_\delta\{D_0|\mathbf{n}\} \right).$$

The risk of the γ-family test averaged of our group size model is denoted $\mathcal{R}_r^{(\gamma)}(\mathbf{d}_{opt})$ and we make equivalent definitions for the ρ-family tests.

Table 6.3 shows results for the error spending method. Values in this table are the average risk for the γ- and ρ-family tests given as percentages of the average risk of the corresponding optimal threshold group sequential test. All the error spending tests have been designed for three groups of 15 observations and have equal type I and type II error in these circumstances, while the optimal tests have these error probabilities averaged over the group size model. The performance of the error spending tests is fair in most cases, although the results for the optimisation of F_{31} indicate that the error spending tests are noticeably sub-optimal in this case. In all cases, and in other examples we have not reported here, the γ-family tests are slightly superior to the ρ-family tests.

To examine how much of the inefficiency shown in table 6.3 is due to the non-optimality of the error spending tests in the fixed groups setting and how much is due to the unanticipated group sizes, we consider the risk of the error spending tests in the fixed groups setting. To do this, we found the fixed groups designs for the sample size sequence $\mathbf{n} = 15, 30, 45)$ with equal type I and type II error probabilities $\alpha = \beta = 0.05$ which minimise each of F_1, F_{21}, F_{31}, and F_4. Table 6.4 shows the risks of the error spending tests given as percentages of the risks of the optimal fixed group procedures; all risks were calculated using the decision cost vector from the relevant optimal threshold group sequential design. In all cases, a large proportion of the inefficiency of the error spending tests comes from the unanticipated group sizes.

Model		γ	ρ
1	1	100.0	100.0
	21	100.3	101.1
	31	102.7	105.4
	4	100.3	101.2
2	1	100.0	100.0
	21	100.3	101.1
	31	102.7	105.5
	4	100.3	101.2
3	1	100.0	100.0
	21	100.3	101.1
	31	102.7	105.5
	4	100.3	101.2

Model		γ	ρ
4	1	100.0	100.1
	21	100.3	101.1
	31	102.9	105.8
	4	100.3	101.2
5	1	100.0	100.0
	21	100.2	100.8
	31	101.7	103.5
	4	100.2	100.8

Table 6.4: *Tabulated values are risks of fixed group sequential tests from the γ-, and ρ-families, designed for the most likely sample size sequence in each model, given as percentages of the risk of the relevant optimal fixed group sequential test. All risks are evaluated using the cost of a wrong decision from the appropriate optimal threshold group sequential test.*

Figure 6-8 compares the risks of the optimal threshold group sequential tests and γ-family tests conditional upon the observed sample size sequence $\mathbf{n}$. The optimal tests are all designed to minimise objective function F_1, the expected sample size when $\mu = \delta/2$, and have average error probabilities $\alpha = \beta = 0.05$. The error spending tests have these error probabilities when the sample sizes follow the planned sequence $\mathbf{n} =$ (15, 30, 45). All risks are calculated using the decision cost vector $\mathbf{d}_{opt}$ which gives the Bayes rule corresponding to the relevant optimal threshold group sequential test; these costs were used in the calculation of tables 6.3 and 6.4. The plotted values are the conditional risks given as percentages of the average risk for the relevant Bayes rule.

The plots in figure 6-8 show a small number of sample size sequences for models 2 and 5 where the conditional risk of the error spending test is slightly lower than for the optimal test. This is because the optimal test minimises the risk over the entire sample size model, and not necessarily for each $\mathbf{n}$. The general pattern in the plots in figure 6-8 is that the points where the conditional risk of the error spending test is closest to

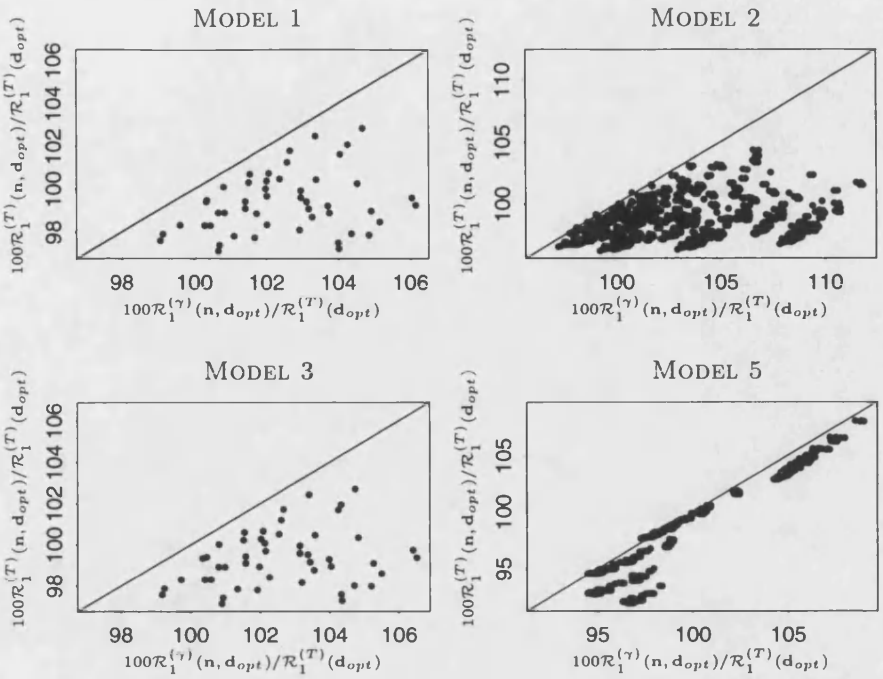


Figure 6-8: *Conditional risks of the optimal threshold group sequential tests and γ-family error spending tests following sample size models 1, 2, 3, and 5. Optimal tests minimise objective function F_1, with average error probabilities $\alpha = \beta = 0.05$. Error spending tests have these error probabilities under the most likely sample size sequence, $\mathbf{n} = (15, 30, 45)$. All risks are given as percentages of the average risk of the relevant optimal threshold group sequential test, and are calculated using the decision costs of this optimal test.*

that of the optimal test represent sample size sequences with smaller maximum sample sizes. When the maximum sample size is considerable larger than the value of N_K for which the error spending test was defined then the conditional risk of the error spending method is further from the conditional risk of the optimal procedure. This indicates that the error spending test does not make efficient use of the observations gathered by overshooting the planned maximum sample size. The plot of conditional risks for tests following model 5 shows a somewhat different pattern. The cluster of points which have the largest conditional risks represent sample size sequences where

$n_K = 40$, that is where we have not seen the number of observations that the error spending test was designed for.

From table 6.3 and figure 6-8, and from similar plots for the other objective functions, we can see that where the average risk for the error spending function is, say, 3% greater then the average risk of the optimal test, then the conditional risks of the error spending test range up to about 6% greater than the corresponding risks of the optimal tests. The cases where the conditional risk of the error spending test is furthest from the conditional risk of the optimal threshold test are when there is a large overshoot of observations and n_K is significantly larger than the planned maximum sample size. Thus, even if the average risk of an error spending procedure is close enough to the risk of the corresponding optimal threshold group sequential test, there may be sample size sequences for which the error spending test performs particularly poorly.

6.4 Discussion

Our optimal threshold group sequential tests provide an alternative means of measuring the performance of the error spending tests to the optimal random group designs we used in chapter 5. The threshold designs provide a more accurate comparison to the circumstances in which an error spending design is often used; when we continue to observe accumulating data until a fixed information level is reached.

The comparisons in §6.3 show that the error spending method can be highly efficient for sample size sequences close to those for which it was designed. However when there is a large overshoot of data, resulting in a maximum sample size considerably larger than that which was planned, the error spending method can be significantly inefficient in the sense of having a much larger expected cost conditional upon the observed sample size than the conditional risk of our optimal threshold group sequential designs. In such circumstances, the error spending method can make little use of the extra information

provided by the large final group of observations as the majority of the type I error has already been spent by this point.

Throughout the results presented in §6.2, we saw that the properties of the optimal threshold group sequential designs were closely linked to the range of sample sizes which could be observed at each analysis. When the range was kept constant, altering the probabilities of given sample size occurring or allowing a greater number of possible sample sizes did little to increase the range of error probabilities conditional upon observed sample size sequences, $\tilde{\alpha}(\mathbf{n})$. Thus, we could design group size models which would allow frequent monitoring of the data and which could approach fully sequential monitoring. So long as the maximum group size of such a model was small, we would maintain an acceptable range of values of $\tilde{\alpha}(\mathbf{n})$ and would gain the reduction of expected sample size associated with frequent monitoring of the data.

Chapter 7

Summary

We have developed group sequential tests which minimise expected sample size for a number of situations and looked at the performance of existing methods relative to the optimal designs. We now wish to summarise the conclusions from each chapter and make some comments on these results in §7.1, before outlining some avenues of further work which remain to be investigated in §7.2. Firstly, we will restate our original problem for convenience; we first defined and discussed this problem in §1.2.1.

We assume a new treatment is being compared to a control, which could be either a placebo or an established treatment and that the difference between the treatment on trial and the control can be measured in some numerical fashion. We denote the treatment difference by μ, with positive values of μ implying superiority of the new treatment. The available data consist of observations $X_1, X_2, \ldots$, where each X_i is independently $N(\mu, \sigma^2)$ distributed with σ^2 known. We wish to test the null hypothesis H_0:$\mu \leq 0$ against the one-sided alternative H_1:$\mu > 0$ with some chosen size α at $\mu = 0$ and with power $1 - \beta$ fixed at $\mu = \delta$. In the trial design stage, δ would be chosen to be some clinically significant and plausible treatment effect.

7.1 Discussion and recommendations

7.1.1 An overview of our results

We have found optimal group sequential tests in several setting and used these tests to gauge the expected sample size properties of the Δ-family and error spending tests. In chapters 3 and 4, we looked at fixed group sequential tests where the sequence of group sizes was known in advance. We considered symmetric designs with equal type I and type II error rates in chapter 3 and designs where the type I and type II error probabilities were not equal in chapter 4. There was very little difference between the results in the symmetric and asymmetric settings. For both sets of fixed groups results we saw that the Δ-family of tests were the least efficient method and that the two families of error spending functions had similar properties, although the γ-family tests were generally slightly more efficient than the ρ-family designs. This pattern continued throughout chapter 5 and while the Δ-family was excluded from chapter 6 the γ-family error spending designs were still slightly superior to the ρ-family tests here. We thus recommend the error spending tests using the γ-family of error spending functions for general use, although care must be exercised in selecting a value of the parameter γ.

7.1.2 Choice of objective functions

We set out to optimise group sequential tests with respect to the expected sample size of the design for various values of μ. These expectations are called objective functions, and we have considered a total of nine objective functions in our discussions. Initially we optimised five objective functions defined by Jennison (1987) and Eales & Jennison (1992).

$$\begin{aligned} F_1 &= \mathbb{E}_{\delta/2}\{N\} \\ F_2 &= \mathbb{E}_0\{N\} = \mathbb{E}_\delta\{N\} \end{aligned}$$

$$
\begin{aligned}
F_3 &= \mathbb{E}_{3\delta/2}\{N\} \\
F_4 &= \frac{1}{5}\mathbb{E}_{\delta/2}\{N\} + \frac{1}{10}\sum_{i=-2, i\neq 2}^{6} \mathbb{E}_{\frac{i}{4}\delta}\{N\} \\
F_5 &= \int \mathbb{E}_{\mu}\{N\}\frac{2}{\delta}\phi\left(\frac{2\mu-\delta}{\delta}\right)\mathrm{d}\mu
\end{aligned}
$$

These objective functions are all symmetric in nature and in chapter 4 we defined six new objective functions to replace F_2 and F_3.

$$
\begin{aligned}
F_{21} &= \frac{1}{2}\left(\mathbb{E}_0\{N\} + \mathbb{E}_\delta\{N\}\right) & F_{22} &= \mathbb{E}_0\{N\} & F_{23} &= \mathbb{E}_\delta\{N\} \\
F_{31} &= \frac{1}{2}\left(\mathbb{E}_{-\delta/2}\{N\} + \mathbb{E}_{3\delta/2}\{N\}\right) & F_{22} &= \mathbb{E}_{-\delta/2}\{N\} & F_{23} &= \mathbb{E}_{3\delta/2}\{N\}
\end{aligned}
$$

Of these, F_{21} and F_{31} are symmetric and designed to specifically replace F_2 and F_3 respectively. Objective functions F_{22}, F_{23}, F_{32} and F_{33} are asymmetric and in §4.3.2, we discussed these objective functions. We recommended that in general use, F_{21} and F_{31} are to be preferred and that the asymmetric objective functions be reserved for certain situations where their asymmetry is beneficial, such as testing a new treatment which is known to have superior secondary characteristics to the control.

Throughout chapter 3, results for the minimisation of F_2, F_4, and F_5 were similar. In chapter 4 we replaced F_2 by F_{21} and here the minimum values for F_{21}, F_4, and F_5 were very close. Moreover, the results presented in figures 3-1 and 4-2 showed that a test optimised for one of these objective functions was close to optimal for the other two objective functions from this trio. The results presented in chapters 5 and 6 did not include the optimising of F_5, but the results for objective functions F_{21} and F_4 were very similar. While we have not presented results in chapters 5 and 6 for the performance of these tests with respect to the objective functions for which they were not optimised, they were similar to those observed in chapters 3 and 4. It seems that in most cases examining results for all of F_{21}, F_4, and F_5 will not be necessary and that it will be sufficient to consider one of these objective functions.

We note that in each chapter the results for F_3 (or F_{31}) have been significantly different from those for the other objective functions considered. Figures 3-1 and 4-2 have shown that tests optimised for F_{31} perform poorly with respect to other objective functions. This is because F_{31} represents very extreme values of μ, and in this situation our optimal action is to encourage very early stopping. Tests optimised for other objective functions are more conservative at early analyses. We also note that of all the tests not optimised for F_{31}, those which minimised F_{21} and F_4 were very nearly equal in their achieved values of F_{31}, and superior to tests optimised for F_1 and F_5.

Based on the results in chapters 3 to 6, we recommend that the efficiency of any group sequential design should be assessed in its performance for F_1 and F_{21}. The objective function F_{31} is a special case and, like the asymmetric objective functions, should be considered only in certain examples. As we noted above, a test which is near-optimal for F_{21} will also have good properties for F_4 and F_5. The computation for tests minimising F_{21} are slightly less intensive than those for F_4 and F_5. Moreover, the monotonicity we desire in finding the optimal Bayes rules (discussed on page 29) can be proven for rules minimising F_1 and F_{21} (Lai, 1973; Brown, Cohen & Strawderman, 1979).

7.1.3 Some comments on the error spending results

The results we have seen in chapters 3 and 4 showed the error spending tests can be highly efficient for equally spaced analyses if the value of γ or ρ is chosen with care. However, they also showed that certain values of the parameters γ and ρ produce tests which are further from optimality or have unacceptably large maximum sample sizes. A prime motivation in our development of the optimal tests in chapters 5 and 6 was to assess the performance of the error spending tests when unanticipated group sizes are observed, the situation for which this method was developed. We have seen that if the number of observations seen is less than that planned, the type II error probability of an error spending design can be significantly inflated over the nominal level. On the

other hand, if more observations become available than was anticipated in the design of the trial then the error spending method can only make use of this to lower the type II error probability, not to improve the type I error rate. In the Bayes setting, this means that the error spending method is further from the optimal expected cost when we see more or less observations than was anticipated.

7.1.4 Practicality of optimal tests with random group sizes

Our optimal random group sequential tests of chapter 5 and optimal threshold group sequential tests of chapter 6 depend upon the specification of a model for the possible sample size sequences which may occur during the conduct of the trial. The comparisons of the results for different group size models in chapter 6 shows that the optimal tests are robust to small changes in the model and hence precise accuracy in the model specification will not be crucial.

In some situations, the specification of a model for the possible sample size sequences will not be possible and in these cases the methods we have developed will not be practical. However, our optimal tests can still play an important role here by assessing the performance of error spending tests or of other flexible designs such as Whitehead's triangular test (Whitehead, 1992, chapter 4).

7.2 Further work

7.2.1 Assessing other test designs

We have used our optimal designs to assess the performance of the Δ-family and error spending methods. These methods were chosen as they are currently widely used and each family incorporates a wide range of tests. However, other group sequential designs exist and it would be interesting to compare their expected sample size properties to

both our optimal tests and the Δ-family and error spending methods. In particular, Whitehead's triangular test (Whitehead, 1992, chapter 4) has been widely used and a software package is available to implement it. We have not considered this method in our discussions as we wish to avoid designs where the maximum sample size is significantly larger than the sample size required by the equivalent non-sequential design; Whitehead's triangular test has a maximum sample size approximately 1.6 times that of the fixed sample test.

It would also be interesting to extend our optimal designs to two-sided tests. Eales & Jennison (1995) and Chang (1996) have used the approach of finding a Bayes decision problem with a solution corresponding to the desired optimal group sequential test in the two-sided setting. Jennison & Turnbull (2000, chapter 2) find two-sided Δ-family tests to be closer to optimality than we have found in the one-sided case, so there are differences in behaviour between these two settings to be investigated.

7.2.2 Sampling schemes

In chapters 5 and 6 we investigated our optimal random groups and threshold designs by considering their properties for a number of models of sample size sequences. These models were chosen to explore the behaviour of our optimal designs and the error spending method, but for the optimal tests to be directly useful models which are closer to the actual accrual of information in a clinical trial would need to be investigated. The independent and identically distributed information increments of chapter 6 would be a suitable place to start, and could lead to very flexible designs.

7.2.3 Variable group size designs

Group sequential designs have been proposed which allow the trial organisers to carry out more frequent analyses if the trial is close to termination at some stage (Lan &

DeMets, 1989). Lan & DeMets found some reduction in expected sample size could be achieved by this method, but that the error rates of the test were perturbed as altering the schedule of analyses in a data dependent fashion violated the sequential design. By building the option of carrying out more frequent analyses into the overall designs of our optimal tests, this perturbation of error rates could be avoided.

This possibility is also attractive from an alternative point of view. We could think in terms of carrying out analyses less frequently if there was little chance of the trial terminating at the next analysis. This would result in a small increase in expected sample size, but would avoid unnecessary meetings of a data monitoring committee and hence ease the logistical burden of carrying out a group sequential clinical trial.

7.2.4 Adaptive allocation

We noted in §1.2.3 that a large amount of research has been done on sequential allocation rules, where new patients are not necessarily evenly divided between the new and control treatments but are allocated in some fashion which is based on the data observed to date. Coad & Rosenberger (1999) have found that combining sequential allocation rules with a fully sequential early stopping rule has potential benefits. In their setting, the response variable was binary with each patient being recorded as a treatment success or failure. By combining a sequential allocation rule with an early stopping rule, the number of treatment failures was reduced. Jennison & Turnbull (2000, chapter 17) have looked at adaptive allocation rules in the group sequential setting, using Δ-family tests, and found that it is possible to significantly reduce the number of patients randomised to the inferior treatment. It would be interesting to investigate the effects of combining a sequential allocation rule with an optimal group sequential design.

7.2.5 Asymptotic normality of test statistics

Throughout this thesis, we have studied the simple problem of independent and identically distributed response data. In §1.2.1 we outlined some of the work that has been done to show that many more sophisticated models have this structure asymptotically. However, it is not clear how well this asymptotic approximation holds in small samples, such as during the early stages of a sequential clinical trial. We have seen that existing group sequential designs can be close to optimal with respect to expected sample size, and it is possible that the inaccuracy induced by the asymptotic nature of some models could be sufficient to offset the gains in efficiency made by the optimal tests.

References

ANDERSON, T.W. (1960). A modification of the sequential probability ratio test to reduce the sample size. *Annals of Mathematical Statistics* **31**, 165–197.

BASU, A., BOAS, A., & GHOSH, J.K. (1991). Sequential design and allocation rules. In *Handbook of Sequential Analysis*, 475–502. New York: Marcel Dekker.

BROWN, L.D., COHEN, A., & STRAWDERMAN, W.E. (1979). Monotonicity of Bayes sequential tests. *Annals of Statistics* **7**, 1222–1230.

CHANG, M.N. (1996). Optimal designs for group sequential clinical trials. *Communications in Statistics - Theory and Methods* **25**, 361–379.

CHANG, M.N., HWANG, I.K., & SHIH, W.J. (1998). Group sequential designs using both type I and type II error probability spending functions. *Communications in Statistics - Theory and Methods* **27**, 1323–1339.

COAD, D.S., & ROSENBERGER, W.F. (1999). A comparison of the randomized play-the-winner rule and the triangular test for clinical trials with binary responses. *Statistics in Medicine* **18**, 761–769.

COOK, R.J. & FAREWELL, V.T. (1996). Incorporating surrogate endpoints into group sequential trials. *Biometrical Journal* **38**, 119–130.

EALES, J.D. (1991). Optimal group sequential tests. *Ph.D. thesis, University of Bath.*

EALES, J.D. & JENNISON, C. (1992). An improved method for deriving optimal one-sided group sequential tests. *Biometrika* **79**, 13–24.

EALES, J.D. & JENNISON, C. (1995). Optimal two-sided group sequential tests. *Sequential analysis* **14**, 273–286.

EMERSON, S.S. & FLEMING, T.R. (1989). Symmetric group sequential test designs. *Biometrics* **45**, 905–923.

FLEMING, T.R. & DEMETS, D.L. (1993). Monitoring of clinical trials: issues and recommendations. *Controlled Clinical Trials* **14**, 183–197.

FOOD AND DRUG ADMINISTRATION (1998). Guidance on statistical principles for clinical trials. *Federal Register* **63:179** 49583–49598.

GELLER, N.L. & POCOCK, S.J. (1987). Interim analyses in randomized clinical trials: ramifications and guidelines for practitioners. *Biometrics* **43**, 213–223.

GHOSH, B.K. (1991). A brief history of sequential analysis. In *Handbook of Sequential Analysis*, 1–20. New York: Marcel Dekker.

HWANG, I.K., SHIH, W.J., & DE CANI, J.S. (1990). Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine* **9**, 1439–1445.

JENNISON, C. (1987). Efficient group sequential tests with unpredictable group sizes. *Biometrika* **74**, 155–165.

JENNISON, C. (1994). Numerical computations for group sequential tests. In *Computing Science and Statistics*, vol. 25. Eds. M. Tarter & M.D. Lock. Interface Foundation of North America. 263–272.

JENNISON, C. & TURNBULL, B.W. (1990). Statistical approaches to interim monitoring of medical trials: a review and commentary. *Statistical Science* **5**, 299–

317.

JENNISON, C. & TURNBULL, B.W. (1991). Group sequential tests and repeated confidence intervals. In *Handbook of Sequential Analysis*, 283–312. New York: Marcel Dekker.

JENNISON, C. & TURNBULL, B.W. (1993). Sequential equivalence testing and repeated confidence intervals, with applications to normal and binary responses. *Biometrics* **49**, 31–43.

JENNISON, C. & TURNBULL, B.W. (1997). Group-sequential analysis incorporating covariate information. *Journal of the American Statistical Association* **92**, 1330–1341.

JENNISON, C. & TURNBULL, B.W. (2000). *Group Sequential Methods with Applications to Clinical Trials.* London: Chapman & Hall.

KIM, K. & DEMETS, D.L. (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* **74**, 149–154.

KORN, E.L. & SIMON, R. (1996). Data monitoring committees and problems of lower-than-expected accrual or event rates. *Controlled Clinical Trials* **17**, 526–535.

LAI, T.L. (1973). Optimal stopping and sequential tests which minimize the maximum expected sample size. *Annals of Statistics* **1** 659–673.

LAN, K.K.G & DEMETS, D.L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–663.

LAN, K.K.G & DEMETS, D.L. (1989). Changing frequency of interim analysis in sequential monitoring. *Biometrics* **45**, 1017–1020.

LAN, K.K.G. & ZUCKER, Z.M. (1993). Sequential monitoring of clinical trials; the role of Brownian motion. *Statistics in Medicine* **12**, 753–765.

LAN, K.K.G., REBOUSSIN, D.M., & DEMETS, D.L. (1994). Information

and information fractions for design and sequential monitoring of clinical trials. *Communications in Statistics - Theory and Methods* **23**, 403–420.

LI, Z & GELLER, N.L. (1991). On the choice of times for data analysis in group sequential clinical trials. *Biometrics* **47**, 745–750.

MCPHERSON, K. (1982). On choosing the number of interim analyses in clinical trials. *Statistics in Medicine* **1** 25–36.

O'BRIEN, P.C. & FLEMING, T.R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.

PAMPALLONA, S. & TSIATIS, A.A. (1994). Group sequential designs for one-sides and two-sided hypothesis testing with provision for early stopping in favour of the null hypothesis. *Journal of Statistical Planning and Inference* **42**, 19–35.

POCOCK, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–199.

POCOCK, S.J. (1982). Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics* **38**, 153–162.

POCOCK, S.J. (1983). *Clinical Trials: a Practical Approach.* Chichester: Wiley.

PROSCHAN, M.A. (1999). Properties of sending function boundaries. *Biometrika* **86**, 466–473.

PROSCHAN, M.A., FOLLMANN, D.A., & WACLAWIW, M.A. (1992). Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics* **48**, 1131–1143.

SCANDINAVIAN SIMVASTATIN SURVIVAL STUDY GROUP (1993). Design and baseline results of the Scandinavian simvastatin survival study of patients with stable angina and/or previous myocardial infarction. *American Journal of Cardiology* **71**, 393–400.

SCHARFSTEIN, D.O., TSIATIS, A.A., & ROBINS, J.M. (1997). Semiparametric efficiency and its implications on the design and analysis of group-sequential studies. *Journal of the American Statistical Association* **92**, 1342–1350.

SLUD, E. & WEI, L.J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association* **77**, 862–868.

WALD, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics* **16**, 117–186.

WANG, S.K. & TSIATIS, A.A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43**, 193–199.

WEISS, L. (1962). On sequential tests which minimize the maximum expected sample size. *Journal of the American Statistical Association* **57**, 551-566.

WHITEHEAD, J (1992). *The design and analysis of sequential clinical trials.* New York: Ellis Horwood.